



NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

DISSERTATION

**VALIDATING COMPUTATIONAL HUMAN BEHAVIOR
MODELS: CONSISTENCY AND ACCURACY ISSUES**

by

Simon R. Goerger

June 2004

Dissertation Supervisor:

Rudolph Darken

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			<i>Form Approved OMB No. 0704-0188</i>	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE June 2004	3. REPORT TYPE AND DATES COVERED Dissertation	
4. TITLE AND SUBTITLE: Validating Computational Human Behavior Models: Consistency and Accuracy Issues			5. FUNDING NUMBERS	
6. AUTHOR(S) Goerger, Simon R.				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000			8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A			10. SPONSORING / MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.			12b. DISTRIBUTION CODE	
13. ABSTRACT (maximum 200 words) As leaders of the Department of Defense (DoD) rely more on modeling and simulation (M&S) to provide information on which they base strategic and tactical decisions, the credibility of simulations becomes more important. This credibility is initially gained through the verification, validation, and accreditation process DoD models are required to undergo prior to their use in simulations. The process of validating behavioral models is not well defined, nor is the process extendable to meet requirements for validating the varied and complex behavioral models. Through a series of empirical studies, this research identifies subject matter expert (SME) biases and their effects on consistency and accuracy of results. This research concludes that a SME's bias has a statistically significant effect on subjective assessment of human performance of urban combat skills. To this end, the research demonstrates how the effects of the natural biases of SMEs can be mitigated based on the scale used to assess assessing human behavior representation (HBR) models, providing a more consistent and accurate means of validating HBR models. In doing so, it assists the DoD M&S Community by providing enhancements to face validation procedures for HBR model implementations for future use in DoD legacy and developmental combat models.				
14. SUBJECT TERMS Validation, Cognitive Model, Modeling and Simulations, Human Behavior Representation, Bias, Multi-Agent Systems, Behavioral Psychology, Cognitive Psychology, VV&A, Human Performance Evaluation			15. NUMBER OF PAGES 340	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL	

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

**VALIDATING COMPUTATIONAL HUMAN BEHAVIOR MODELS:
CONSISTENCY AND ACCURACY ISSUES**

Simon R. Goerger
Major, United States Army
B.S., United States Military Academy, 1988
M.S., Naval Postgraduate School, 1998

Submitted in partial fulfillment of the
requirements for the degree of

**DOCTOR OF PHILOSOPHY IN MODELING, VIRTUAL ENVIRONMENTS
AND SIMULATION (MOVES)**

from the

**NAVAL POSTGRADUATE SCHOOL
June 2004**

Author:

Simon R. Goerger

Approved by:

Dr. Rudolph P. Darken
Chair, MOVES Academic Committee
Dissertation Supervisor

COL Michael L. McGinnis
Professor and Head
Department of Systems Engineering
West Point, New York

Dr. Michael J. Zyda
Director of MOVES

Dr. Nita L. Miller
Professor of Operations Research

Dr. Christian Darken
Professor of Computer Science

Prof Susan G. Hutchins
Professor of Information Sciences

Approved by:

Dr. Rudolph P. Darken, Chair, MOVES Academic Committee

Approved by:

Julie Filizetti, Associate Provost for Academic Affairs

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

As leaders of the Department of Defense (DoD) rely more on modeling and simulation (M&S) to provide information on which they base strategic and tactical decisions, the credibility of simulations becomes more important. This credibility is initially gained through the verification, validation, and accreditation process DoD models are required to undergo prior to their use in simulations. The process of validating behavioral models is not well defined, nor is the process extendable to meet requirements for validating the varied and complex behavioral models. Through a series of empirical studies, this research identifies subject matter expert (SME) biases and their effects on consistency and accuracy of results. This research concludes that a SME's bias has a statistically significant effect on subjective assessment of human performance of urban combat skills. To this end, the research demonstrates how the effects of the natural biases of SMEs can be mitigated based on the scale used to assess human behavior representation (HBR) models, providing a more consistent and accurate means of validating HBR models. In doing so, it assists the DoD M&S Community by providing enhancements to face validation procedures for assessing HBR model implementations for future use in DoD legacy and developmental combat models.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

I.	INTRODUCTION.....	1
A.	THESIS STATEMENT	1
B.	PROBLEM STATEMENT AND APPROACH.....	2
C.	GOAL.....	4
D.	SIGNIFICANT CONTRIBUTIONS.....	4
E.	DISSERTATION ORGANIZATION	5
II.	BACKGROUND AND RELATED WORK	7
A.	VERIFICATION, VALIDATION AND ACCREDITATION	8
B.	PSYCHOLOGY	15
C.	COGNITIVE MODELS	20
1.	Representations and Architectures	20
2.	Implementations.....	23
D.	HUMAN BEHAVIOR REPRESENTATION	26
1.	Human Behavior Representation Verification and Validation Procedures	26
2.	Referent Categories	28
3.	Face Validation.....	30
4.	Subject Matter Experts	30
5.	Issues	31
E.	VALIDATION EFFORTS OF HUMAN BEHAVIOR MODELS	32
F.	HUMAN PERFORMANCE EVALUATION	34
1.	Procedural Versus Declarative Knowledge	34
2.	Bias	35
3.	Performance Appraisal	36
G.	NATURALISTIC DECISION-MAKING	38
H.	ASSESSMENT OF PREVIOUS WORK.....	42
III.	METHODOLOGY AND EXPERIMENTAL DESIGN.....	49
A.	SCOPE	52
B.	EXPERIMENTAL DESIGN.....	53
1.	Study Simulation.....	56
2.	Simulation Environment	57
3.	Scenarios	60
4.	Data Collection Techniques	62
C.	STUDY #1 EXPERIMENTAL DESIGN	64
1.	Hypotheses.....	65
2.	Design	66
3.	Procedures	68
4.	Set-Up.....	69
5.	Study Phases	71
D.	STUDY #2 EXPERIMENTAL DESIGN	72
1.	Hypotheses	72
2.	Design	74
3.	Procedures, Set-Up, and Study Phases	76

IV.	RESULTS	79
A.	GENERAL.....	79
1.	Subject Matter Expert Demographics	79
2.	Data Set.....	84
B.	BIAS PATTERNS.....	85
C.	ASSESSMENT	93
D.	STEPWISE LOGISTICS REGRESSION.....	113
V.	DISCUSSION	119
A.	SIMULATION BELIEF.....	119
B.	ASSESSMENT SCALES	119
C.	BIAS	121
D.	VALIDATION CRITERIA.....	122
E.	BIAS AND PERSONALITY.....	125
VI.	CONCLUSIONS, SUMMARY, AND RESEARCH AGENDA.....	129
A.	CONCLUSIONS	129
B.	SUMMARY	130
C.	FACE VALIDATION PROCESS RECOMMENDATIONS	134
D.	TOWARDS A RESEARCH AGENDA	137
1.	Issues with Human Behavior Representation Models.....	138
2.	Validation Issues	140
	APPENDIX A. REFERENT FOR HUMAN BEHAVIOR REPRESENTATION MODELS	147
	APPENDIX B. EXPERIMENTAL PROCEDURES	151
	APPENDIX C. PARTICIPANT TASKS	155
	APPENDIX D. ASSESSMENT OF PARTICIPANT TASKS	159
	APPENDIX E. EXPERIMENT MATERIALS.....	163
	APPENDIX F. STUDY ENVIRONMENTS AND SCENARIOS.....	167
	APPENDIX G. PARTICIPANT DEMOGRAPHICS, EXPERIENCE, AND TRAINING QUESTIONNAIRE	195
	APPENDIX H. CONSENT FORMS	199
	APPENDIX I. DEBRIEFING HANDOUT.....	203
	APPENDIX J. EXPERIMENT EXIT QUESTIONNAIRE	205
	APPENDIX K. ASSESSMENT WORKSHEETS.....	209
	APPENDIX L. BRIEFING SCRIPTS	225
	APPENDIX M. EXPERIMENT BRIEFING SLIDES.....	237
	APPENDIX N. SUPPORTING FIGURES AND TABLES FOR DATA ANALYSIS...251	
	APPENDIX O. ILLUSTRATION OF FACE VALIDATION SHORTCOMINGS	263
	APPENDIX P. KEY PLAYERS IN VERIFICATION, VALIDATION AND ACCREDITATION	269

APPENDIX Q. VALIDATION PLAN.....	271
APPENDIX R. MODEL TAXONOMIES	285
GLOSSARY.....	293
LIST OF REFERENCES	301
INITIAL DISTRIBUTION LIST	315

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

Figure 1.	Thesis Objective: To Define the Common Area.....	1
Figure 2.	Research Focus and Contributions to Face Validation Process After [6].....	5
Figure 3.	DoD Modeling and Simulation Landscape.....	8
Figure 4.	Modeling and Simulation Problem Solving Process From [DEPA 01e].....	9
Figure 5.	Birta and Özmırak Validation/Verification After [BIRT 96]	11
Figure 6.	Essential Steps for Validating Models and Simulations From [DEPA 00b] ...	14
Figure 7.	Waugh and Norman’s Model of Human Memory From [SOLS 01].....	19
Figure 8.	Verification, Validation, and Accreditation Tasks for a Human Behavior Representation Model After [23].....	27
Figure 9.	Metamodel Correspondence From [CAUG 95].....	44
Figure 10.	Global Architecture for Birta and Özmırak’s Automated Result Validation Model From [BIRT 96].....	45
Figure 11.	Area of Interest for Validating Human Behavior Representations Models.....	49
Figure 12.	Model and Simulation Classifications From [45].....	57
Figure 13.	Training and Practice Environment Sketch	58
Figure 14.	McKenna Test Environment Sketch From [STAT 03].....	59
Figure 15.	McKenna Test Environment Defensive Sketch.....	60
Figure 16.	Training and Practice Environment Model Display and Scenario.....	61
Figure 17.	MANA Display of McKenna Test Environment, Offensive Scenario #2	62
Figure 18.	Room Layout for September Data Collection	70
Figure 19.	Room Setup for September Data Collection.....	71
Figure 20.	Room Layout for October Data Collection.....	76
Figure 21.	Room Setup for October Data Collection.....	77
Figure 22.	Experiment Assessment Sublevels and Levels	85
Figure 23.	Bias Manifestations.....	86
Figure 24.	Performance Bias Example.....	87
Figure 25.	Anchoring Bias Examples.....	88
Figure 26.	Contrast Bias Example.....	90
Figure 27.	Confirmation Bias Examples	92
Figure 28.	Study #1, Subject Matter Expert Bias for 7-Point Likert Scale.....	93
Figure 29.	Distribution of Subject Matter Experts’ Normalized Responses to Question Overall 1	95
Figure 30.	Subject Matter Expert Normalized Responses to Subtask 2, Task 1, Scenario 1.....	96
Figure 31.	Subject Matter Expert Normalized Responses to Overall 1	97
Figure 32.	Intra-SME Mean Consistency Scores	101
Figure 33.	Intra-SME Subtask-to-Task Consistency Scores	102
Figure 34.	Intra-SME Mean Consistency Impact Scores	105
Figure 35.	Intra-SME Mean Accuracy Scores	108
Figure 36.	Intra-SME Mean Accuracy Impact Scores	111
Figure 37.	Distribution of Subject Matter Experts’ Normalized Responses to Question Overall 1 without Biased Subject Matter Expert Responses.....	112
Figure F1.	Training and Practice Environment Building Numbers.....	169
Figure F2.	Training and Practice Environment Terrain Sketch.....	170

Figure F3.	Training and Practice Environment Floor Plan (Display Sketch).	171
Figure F4.	Offensive Test Environment Sketch From [STAT 03].....	172
Figure F5.	Offensive Test Environment Building Numbers	173
Figure F6.	Offensive Test Environment Terrain Sketch.	174
Figure F7.	Defensive Test Environment Floor Plan (Display Sketch).....	175
Figure F8.	Defensive Test Environment Terrain Sketch.....	176
Figure F9.	Training and Practice Environment Sketch Symbolology From [DEPA 97a] .	178
Figure F10.	Mines and Wire Obstacles From [US L 99] [SOLD 03]	179
Figure F11.	Weapons Systems From [US L 99]	179
Figure F12.	Training and Practice Environment Floor Plans with Example Defensive Positions for Building EI	180
Figure F13.	MANA Display of Training and Practice Environment; Initial State for Defensive of Building AI	181
Figure F14.	MANA Display of Training and Practice Environment; Final State and Blue-Force Routes for Defensive of Building AI	182
Figure F15.	McKenna Offensive Scenario Templated Enemy After [STAT 03]	184
Figure F16.	McKenna Offensive Scenario Blue-Force Graphics.....	185
Figure F17.	Offensive Sketch Symbolology From [DEPA 97a]	186
Figure F18.	MANA Display of McKenna Offensive Scenario #1; Initial State	187
Figure F19.	Natick Study; Movement Scenario 1 From [STAT 03].....	188
Figure F20.	MANA Display of McKenna Offensive Scenario #1; End State, Blue Force Routes, and Engagement Locations.....	188
Figure F21.	MANA Display of McKenna Offensive Scenario #2; Initial State	189
Figure F22.	MANA Display of McKenna Offensive Scenario #2; End State, Blue Routes, and Engagement Locations.....	190
Figure F23.	MANA Display of McKenna Defensive Scenario; Initial State	191
Figure F24.	MANA Display of McKenna Defensive Scenario; End State, Blue Routes, and Engagement Locations for Squad Leader	192
Figure F25.	MANA Display of McKenna Defensive Scenario; End State, Blue Routes, and Engagement Locations for Team 1	193
Figure N.1.	Participant NEO-FFI Raw Score Fitted Normal Quad Chart	254
Figure N.2.	Participant NEO-FFI Raw Score Fitted Normal Chart - Conscientiousness .	254
Figure N.3.	Intra-SME Task-to-Scenario Consistency Scores.....	256
Figure N.4.	Intra-SME Scenario-to-Overall Consistency Scores	257
Figure N.5.	Assessment Accuracy: Level	258
Figure N.6.	Assessment Accuracy: Simulation by Scale	259
Figure O.1.	Templated Enemy Situation.....	263
Figure O.2.	Friendly Course of Action	264
Figure O.3.	Movement of Forces	265
Figure O.4.	Enemy Action	266
Figure O.5.	Friendly Forces Potential Decisions	267
Figure R.1.	Simulation Typologies From [HUGH 97] [GOER 03].....	286
Figure R.2.	Army Modeling and Simulation Office's Hierarchy of Modeling and Simulation From [AMSO 00]	288
Figure R.3.	Hierarchy of Modeling and Simulation From [107]	289
Figure R.4.	Model and Simulation Taxonomies From [109].....	292

LIST OF TABLES

Table 1.	Steps in Verification and Validation Process Where Comparison Techniques Best Apply After [DEPA 00b]	12
Table 2.	General Limitations of Different Comparison Techniques From [DEPA 00b]	12
Table 3.	Requirements and Sources of Validation Information From [DEPA 00b]	15
Table 4.	Model Architecture Action Information Sources After [PEW 98] [OSBO 02]	23
Table 5.	Comparison of the Validation of Different HBRs From [PEW 98]	33
Table 6.	Study #1, Primary Factor Levels	67
Table 7.	Study #1, Experimental Layout by Subject Matter Expert Group	68
Table 8.	Study #2, Primary Factor Levels	75
Table 9.	Experimental Layout for Study #2	75
Table 10.	Subject Matter Expert Demographics: Time in Service Data	80
Table 11.	Subject Matter Expert Demographics: Deployment Data	81
Table 12.	Subject Matter Expert Demographics: Game and Model Experience	82
Table 13.	Data Set Groupings	84
Table 14.	Mean Values for Normalized, Overall Assessment Scores	94
Table 15.	Ordinal Logistical Fit for Normalized Assessment Values	98
Table 16.	Ordinal Logistical Fit for Normalized Consistency Scores	100
Table 17.	Ordinal Logistical Fit for Normalized Consistency Impact Scores	104
Table 18.	Ordinal Logistical Fit for Normalized, Absolute Value, Accuracy Scores ..	107
Table 19.	Ordinal Logistical Fit for Normalized Accuracy Impact Scores	110
Table 20.	Normalized, Mean Overall Assessment Scores - Minus Bias	112
Table 21.	Term Estimates from Ordinal Logistic Fit for Accuracy Impact Scores	114
Table N.1.	Participant Demographics: Education & Service Data	251
Table N.2.	NEO-FFI Raw Score Conversions From [COST 92]	252
Table N.3.	Participant NEO-FFI Raw Score Statistics	253
Table N.4.	Likelihood Ratio Tests: Subtask-to-Task Effect - Consistency	255
Table N.5.	Consistency Means of Normalized Values: Subtask-to-Task	255
Table N.6.	Likelihood Ratio Tests, Task-to-Scenario Effect - Consistency	255
Table N.7.	Consistency Means of Normalized Absolute Values: Task-to-Scenario	256
Table N.8.	Consistency Means of Normalized Values: Task-to-Scenario	257
Table N.9.	Likelihood Ratio Tests, Scenario-to-Overall Effect - Consistency	257
Table N.10.	Ordinal Logistical Fit for Normalized Accuracy Scores	258
Table N.11.	Ordinal Logistical Fit for Normalized Accuracy Impact Scores	259
Table N.12.	Normalized, Mean Overall Assessment Scores - Minus Performance Bias ..	260
Table N.13.	Normalized, Mean Overall Assessment Scores - Minus Anchoring Bias	260
Table N.14.	Normalized, Mean Overall Assessment Scores - Minus Contrast Bias	260
Table N.15.	Normalized, Mean Overall Assessment Scores - Minus Confirmation Bias ..	261
Table N.16.	Normalized, Mean Accuracy Impact Score Difference – Results Without Bias Minus All Results	261

Table N.17.	Percentage Change in Normalized, Mean Accuracy Impact Score Difference – Results Without Bias Minus All Results	261
Table P.1.	Typical Roles and Responsibilities Associated with Modeling and Simulation Verification, Validation and Accreditation From [DEPA 01e]...	269

LIST OF EQUATIONS

Equation 1.	Sublevel Consistency Score.....	99
Equation 2.	Level Consistency Score.....	99
Equation 3.	Mean Sublevel Consistency Score.....	103
Equation 4.	Binned Sublevel Consistency Score	103
Equation 5.	Binned Level Consistency Score	103
Equation 6.	Sublevel Consistency Impact Score.....	103
Equation 7.	Level Consistency Impact Score.....	103
Equation 8.	Sublevel Accuracy Score	106
Equation 9.	Level Accuracy Score	106
Equation 10.	Normalized Element Score	109
Equation 11.	Binned Element Score.....	109
Equation 12.	Binned Assessment Key Score	109
Equation 13.	Sublevel Accuracy Impact Score	109
Equation 14.	Level Accuracy Impact Score.....	109

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ABBREVIATIONS, ACRONYMS, SYMBOLS

1LT	First Lieutenant
ACTA	Applied Cognitive Task Analysis
ACT-R	Adaptive Control of Thought
AFRL	Air Force Research Laboratory
AI	Artificial Intelligence
AL	Artificial Life
AMBR	Agent-based Modeling and Behavior Representation
AMSO	Army Model and Simulation Office
ARTEP	Army Training and Evaluation Program
AT-4	84mm Rocket Launcher (M136)
ATM	Anti-Tank Mine
ATP	Ant-Personnel Mine
BARS	Battlefield Augmented Reality System
C4I	Command, Control, Communications, Computers, and Intelligence
CAS	Complex Adaptive System
CASTFOREM	Combined Arms and Support Task Force Evaluation Model
CATD	Combined Arms & Tactics Directorate
CDM	Critical Decision Method
CGFs	Computer Generated Forces
CMAS	Connector-Based Multi-Agent System
CMTDD	Course Management Training Development Division
COGNET	COGnition as a NETwork of Tasks
COMBAT ^{XXI}	Combined Arms Analysis Tool for the XXI st Century
COP	Common Operating Picture
CPT	Captain
CPV	Cognitive Process Validation
CTA	Cognitive Task Analysis
DBBL	Dismounted Battlespace Battle Lab
D-COG	Distributed Cognition (AFRL's agent-based modeling architecture)
DMSO	Defense Modeling and Simulation Office
DNA	Decompose, Network, and Assess
DoD	Department of Defense
DOT	Directorate of Operations and Training
DOTSE	Defence Operational Technology Support Establishment (New Zealand)
DSS	Decision Support System
EPIC	Executive-Process Interaction Control
FDC	Fire Direction Center
FDR	Functional Design Requirement
FO	Forward Observer
GOMS	Goals, Operators, Methods, and Selection Rules

HBR	Human Behavior Representation
HBTWG	Human Behavior Technology Working Group
ICCC	Infantry Captains Career Course
ISAAC	Irreducible Semi-Autonomous Adaptive Combat
JCATS	Joint Conflict And Tactical Simulation
JRTC	Joint Readiness Training Center, Fort Polk, LA
JSAF	Joint Semi-Autonomous Forces (Simulation)
JWARS	Joint Warfare System
M&S	Modeling and Simulation
M136	84mm Anti-Armor Rocket Launcher (AT-4)
M14	Anti-Personnel Mine
M15	Anti-Tank Mine
M16	Anti-Personnel Mine
M16A2	Assault Rifle
M18A1	Claymore Anti-Personnel Mine
M19	Anti-Tank Mine
M203	Grenade Launcher, attached to M16A2
M21	Anti-Tank Mine
M24	Sniper Rifle
M249	Squad Automatic Weapon (SAW) – Light Machine Gun
MAJ	Major
MANA	Map Aware Non-uniform Automata
MAS	Multi-Agent System
MCCDC	Marine Corps Combat Development Command
MI	Military Intelligence
MOE	Measure of Effectiveness
MOP	Measure of Performance
MOUT	Military Operations in Urban Terrain
MTP	Mission Training Plan
NAVMSMO	Navy Modeling & Simulation Management Office
NDM	Naturalistic Decision-Making
NEO-FFI	Neuroticism, Extraversion, and Openness Five-Factor Inventory
NPS	Naval Postgraduate School
NTC	National Training Center, Fort Erwin, CA
OBJ	Objective
OneSAF	One Semi-Automated Force
OOTW	Operations Other Than War
PARI	Prediction, Action, Result, Interpretation
PDC	Primary Data Collector
PEO STRI	Program Executive Office for Simulation, Training, and Instrumentation
PI	Primary Investigator
PO2	Petty Officer Second Class
POW	Prisoner of War
QRF	Quick Reaction Force

RA	Research Assistant
ROE	Rules of Engagement
RPD	Recognition-Primed Decision model
RPG	Recommended Practices Guide (DMSO V&V TWG)
SA	Situational Awareness
SAIC	Science Applications International Corporation
SAW	Squad Automatic Weapon – Light Machine Gun (5.56mm)
SMART	Simulation and Modeling for Acquisition, Requirements, and Training
SME	Subject Matter Expert
Soar	State, Operator and Result (Model)
TADMUS	Tactical Decision-making Under Stress
TRAC	Training and Doctrine Command Analysis Center
TTPs	Tactics, Techniques, and Procedures
TWG	Technical Working Group
UML	Unified Modeling Language
V&V	Verification and Validation
VKB	Validation Knowledge Base
VV&A	Verification, Validation and Accreditation
WP	White Phosphorous
WSMR	White Sands Missile Range (TRAC)
XML	Extensible Mark-up Language

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

As with any endeavor of magnitude, many persons played a role in this project, from inception to completion.

First, I thank God, my wife, Niki, and our son Thomas Gabriel. Through their love I was upheld, guided, inspired, and nourished in mind, body, and soul every step of the way. Along with my family and in-laws, they provided the foundation for my daily efforts through encouragement and prayers.

Next, I am indebted to Rudy Darken, who led, prodded, and challenged me, readying me for each new phase of the endeavor.

COL McGinnis, as coach, visionary, and mentor, refined my skills and focus. Mike Zyda, Nita Miller, Sue Hutchins, and Chris Darken shared experience and wisdom. John Hiles, LTC Gene Paulo, LCDR Russ Schilling, and Elaine Schilling lent early assistance in defining the scope of my work.

My fellow doctoral students, Curt Blais, Joerg Wellbrink, David Wells, Dietmar Kunde, and Perry McDowell, were ever on hand with camaraderie, insight, and a critical eye. Thanks to the MOVES staff, especially Margaret Davis, John Falby, Brigitte Kirchenbauer, and Cecelia Childers, for cheerful handling of the paper work.

The individuals to whom I am indebted represent a number of agencies. The Defense Modeling and Simulations Office, under the supervision of Technical Director for Verification, Validation and Accreditation (VV&A) Simone Youngblood, kept me motivated and reviewed the work in progress. Invaluable assistance in documentation and procedures review to ensure soundness in the validation process was provided by Scott Harmon of ZETETIX and the DMSO VV&A Technical Working Group; Dale Pace from the Johns Hopkins University Applied Physics Laboratory; Susan Solick, the Army M&S VV&A Standards Category coordinator; Marcy Stutzman from Northrop Grumman of the Navy VV&A team; Jennifer Park, the Navy Modeling and Simulation Management Office VV&A leader; and Bill Rostad.

Without the scores of subjects who participated in this research, this work could never have been. Special gratitude is extended to MAJ William Bestermann, MAJ Mark

Hollis, and MAJ John Best of the Infantry Officers Career Course at Fort Benning, GA, who coordinated for and supervised the 182 students who served as subject-matter experts. Two other men played a major role in data collection: my research assistants, CPT Benjamin Tipton, US Army, and Petty Officer Second Class Brian Wood, US Navy. I also thank the officers and staff of the Training and Doctrine Command Analysis Center, Monterey, who provided subject-matter experts for the pilot study.

Finally, I wish to recognize three organizations that supplied vital resources: the Marine Corps Combat Development Command allowed me to use the agent-based model MANA; Natick Soldier Center offered human-performance data of soldiers conducting operations in the McKenna, MOUT Site at Fort Benning; and the Navy Modeling and Simulation Management Office financed equipment and travel in conjunction with dissertation research.

To all, my deepest gratitude.

I. INTRODUCTION

A. THESIS STATEMENT

The representation of human behaviors in computer simulations is a relatively new and very complex area of research that lies at the nexus of modeling and simulations, and behavioral and cognitive psychology. Researchers in this area attempt to model human behavior and simulate human performance using computer simulations primarily developed and used for training, analysis, and research. Each community approaches modeling human behavior from a different direction. This research identifies the boundaries of the area that are common to the three domains, and presents a new methodology for validating models with embedded human behavior representation. The experimental application of the validation methodology to two data sets helped sharpen the focus of the boundaries where the three domain areas intersect as shown in Figure 1.

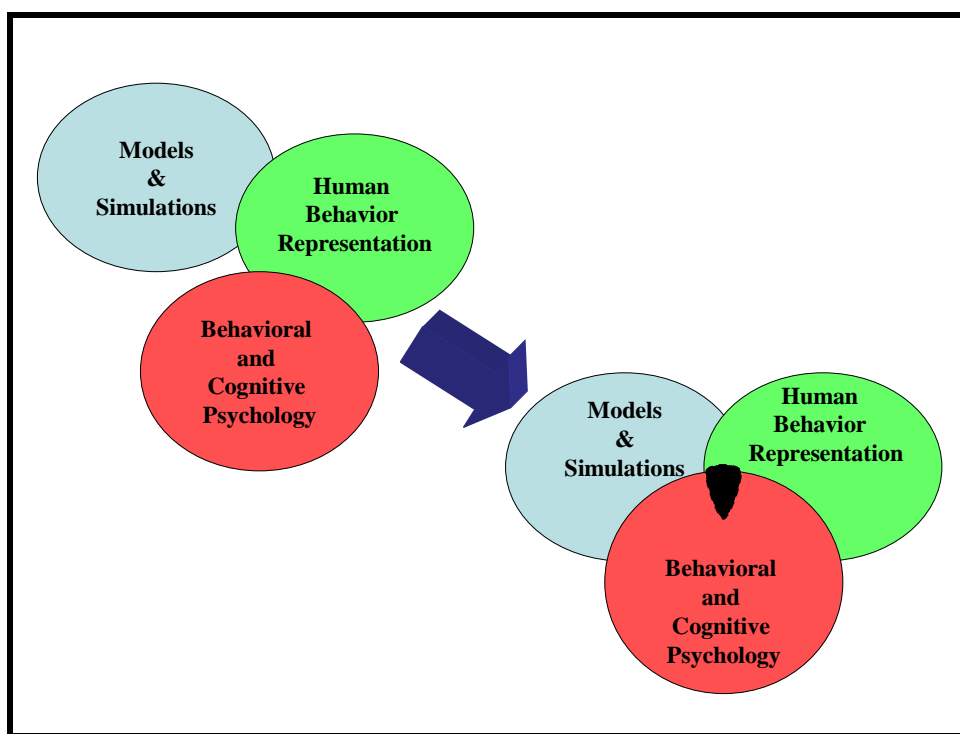


Figure 1. Thesis Objective: To Define the Common Area

This dissertation research studies how to use subject matter experts to evaluate models of human behavior in order to obtain consistency and accuracy in the validation

process. Through a series of empirical studies, results show evaluator biases are a significant contributor to error and that we can mitigate the influence of biases through the use of different assessment scales and by providing appropriate feedback during the validation process.

B. PROBLEM STATEMENT AND APPROACH

Department of Defense (DoD) models and simulations must continually pursue new capabilities to meet the training and analytical needs of America's military establishment. As researchers have improved the fidelity of physics-based models, expectations have risen among model users, analysts, and decision makers to improve how models represent human behaviors. As the capabilities of models increase, so do their complexities. The introduction of complex behavior models into the M&S domain, and the lack of verified supporting data, have made the process of validating models more difficult.

The process of validating physics-based models is well-defined using long-established standards. However, as the background research will show, the process of validating behavioral models is not as well-defined. The validation process developed, matured, and refined over time for physics-based models is not well suited for validating behavioral models. This is due to several factors:

- The nondeterministic nature of human cognitive processes [DEPA 01f];¹
- The large set of interdependent variables making it impossible to account for all possible interactions [DEPA 01f];
- Inadequate metrics for validating HBR models;
- The lack of a robust set of environmental data to run behavioral models for model validation; and
- No uniform, standard method of validating cognitive models.²

It is the contention put forward in this dissertation that subject matter expert (SME) bias demonstrated in the assessment of human behavior representations for human

¹ Human behavior is nondeterministic or appears to be due to numerous influencing variables. The nonlinear relationship between input variables and resulting behaviors makes it impossible to predict the validity of behaviors for one set of inputs based on the validity of behaviors resulting from a different set of inputs [DEPA 01f].

² Cognitive models "describe the detection, storage, and use of information" [SOLS 01]. This refers to models that simulate the human thought process to select actions for execution during a simulation.

ground combatants can be identified, measured, and mitigated using techniques and standards similar to what is used in assessing the performance of actual soldiers.³ We test this hypothesis using a series of studies of company grade Army officers that analyzes their assessment of the performance of soldier tasks derived from *ARTEP 7-8-MTP: Mission Training Plan for the Infantry Rifle Platoon and Squad* [ARTE 01], and performed in a simulated virtual environment.

During experimentation sessions, SMEs quantitatively assessed the degree to which computer objects representing soldiers performed tasks to standard. The approach demonstrates some of the strengths and weaknesses of the methodology for assessing computerized human behavioral models.

Human behavior of interest to the military community occurs in complex, multi-dimensional environments with an abundance of stimuli and in a time, space continuum. Therefore, the environments and scenarios developed for studying human behavior models must also reflect these complexities. Given this context of the problem, we propose two major assumptions that bound the scope of the dissertation. First, based on computational requirements alone, it is beyond the limits of current technology to develop a computable mathematical algorithm or computer program to assess nondeterministic, nonlinear human behavior.⁴ Second, in the same way that the study of naturalistic decision-making asserts that fully understanding human behavior in other than the environments and situations where it naturally occurs is virtually meaningless, validating models of human behavior outside the context of the environment is also meaningless.⁵

³ The term subject matter expert (SME) is used throughout this document referring to raters and study participants. Although not meeting all the requirements specified by Klein, the SMEs used in the dissertation are experienced individuals in the area of military operations in urban terrain.

⁴ This is a version of the halting problem and is based on Turing's Theory of Computability. The halting problem is an attempt to determine if an algorithm will run to completion given a set of inputs. Given an algorithm that produces nondeterministic, nonlinear behavior, one cannot write an algorithm to assess its performance [BLAC 03]. This is because the behavior algorithm may produce an unexpected result which the test algorithm would not understand as an end state and therefore, the test algorithm would continue to execute in a do loop, never ending. Thus, since we cannot use a computer algorithm to assess the validity of a nondeterministic, nonlinear human behavior model, and we are left with the use of subject matter experts to do so.

⁵ Naturalistic decision-making is "the study of how people use their experience to make decisions in field settings" [KLEI 01].

C. GOAL

The intended outcome of *any* validation process applied to models of human behavior is to assure *simulated* human behavior is consistent with *actual* human behavior under the constraints and context of a specific domain. The overarching goal of this dissertation, therefore, is to develop a methodology for validating HBR model implementations for use in Department of Defense training and research models and simulations. In accomplishing this goal, we identify and mitigate issues regarding validation and use of HBR models implemented in legacy and emergent combat simulations.

D. SIGNIFICANT CONTRIBUTIONS

The primary scientific advancement of the research addresses how consistently and accurately SMEs validate real or simulated human performance. The consistency and accuracy of SME assessments of HBR models directly impacts model consistency and accuracy and consequently, what we know about how it will perform in novel situations. The research demonstrates the effect of personality, bias, and assessment scale on the consistency and accuracy of SME responses during the validation process. It provides a means of identify SME bias which can then be mitigated through training or use of human performance evaluation techniques. The results make it possible to provide a more consistent and accurate assessment of the HBR model providing the M&S community with better models for training and analysis.

A second major contribution of the dissertation methodology is identifying the boundaries of the common area between the three communities that will be brought together for the validation of human based models. This work lays the foundation for the research agenda designed to improve the process of validating human behavior representation models. Figure 2 depicts the contributions to the face validation process. Other notable contributions are as follows:

- Lessons learned from the use of human behavior evaluation techniques in the assessment of human behavior models;
- Identifies means to increase the consistency and accuracy of ‘face validation’ procedures for HBR models (M&S);

- Formulates new techniques for identifying and measuring the presence and impact of SME consistency and accuracy (M&S);
- Identifies quantitative patterns of bias based on SME responses to assessment questions (M&S & Psychology);
- Identifies methods for removal of SME bias to mitigate SME inconsistencies and inaccuracies (M&S & Psychology);
- Establishes a statistically significant relationship between bias and Neuroticism, Extraversion, and Openness Five-Factor Inventory personality styles (M&S & Psychology); and
- Proposes a research agenda for the future enhancement of human behavior representation model validation procedures (DoD M&S Community).

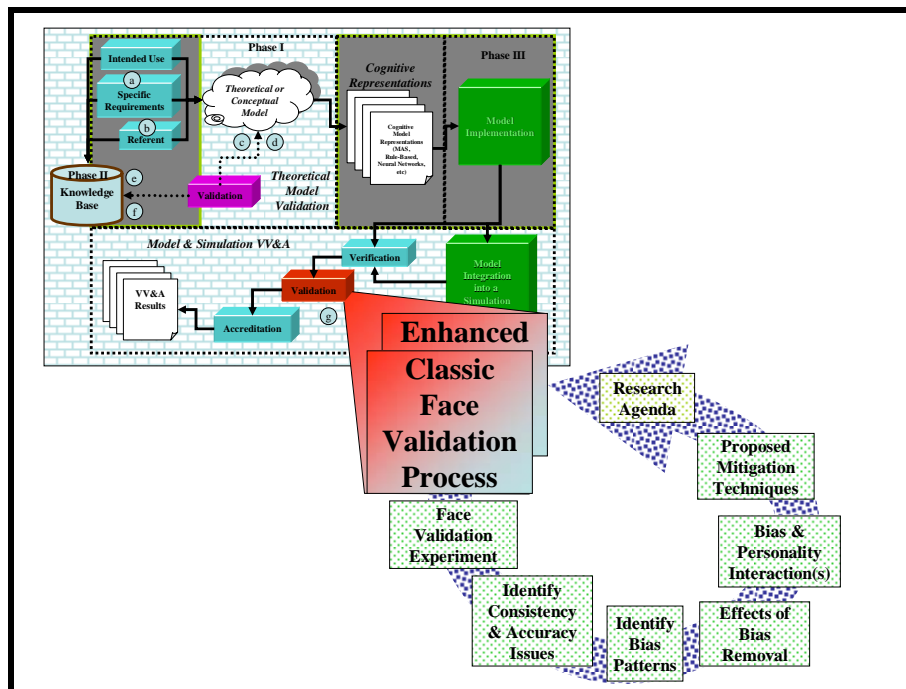


Figure 2. Research Focus and Contributions to Face Validation Process After [6]

E. DISSERTATION ORGANIZATION

The remainder of the dissertation is organized as follows. Chapter II provides a detailed description of previous work, current research, and background material relevant to this dissertation. Chapter III conceptualizes the validation methodology and describes the design of experiments. Analysis of the data collected from the experiments and studies is presented in Chapter IV. This chapter also describes the types of bias and

⁶ See DEPA 00b, DEPA 01e, and DEPA 01f.

defines data consistency and accuracy. Chapter V discusses data analysis, results, and presents solutions to the research questions. The dissertation concludes with Chapter VI which summarizes results, provides recommended face validation procedures, and outlines a research agenda for dealing with issues such as face validation of HBR models. The dissertation's appendices provide additional material from the research studies and validation process for future reference, as well as a glossary of terms, list of references, and a distribution list.

II. BACKGROUND AND RELATED WORK

Traditionally, most DoD models and simulations of military forces have focused on replicating armed conflict between two or more sides. This paradigm of physics-based, force-on-force models relies on mathematical algorithms instantiated in computer programs to study battle damage aspects of combat. Metrics, such as the probability of hits and kills, are used to assess the effectiveness of various weapon systems and munitions, fired from various platforms, subject to specific environmental conditions and target types. Over the past decade, however, military operations have placed more emphasis on the actions of the participants rather than on the characteristics of the weapon systems. In response to this new focus, M&S research has shifted to the development of models that represent the human dimensions of operations other than war (OOTW) and combat operations.

As stated in Chapter I, the goal of this research is to integrate into a single framework, a new methodology for validating human behavior models that draws upon three distinct domains: entity-level combat simulations, human behavior representation, and cognitive psychology.⁷ The body of behavioral research encompasses many elements of human decision-making to include information gathering, situational awareness, and information processing and communicating. Cognitive models attempt to replicate the human decision-making process through models of human behavior. Cognitive models, linked with physics models, attempt to reproduce human behaviors in a dynamic, simulated environment.

Most behavioral models today deal with a very narrow range of human behaviors that are generally categorized as reactive or procedural. Reactive models follow an input-output, cause and effect protocol where a simulated ‘human’ agent executes an action that responds to a stimulus injected into the current situation. Procedural models require simulation agent to follow a prescribed protocol for analyzing a situation, processing information, selecting an appropriate action, and then executing the action. Within the

⁷ These domains use numerous terms interchangeably. To reduce confusion and to ensure this research conveys its points, we define some terms in footnotes. The Glossary at the back of the dissertation contains a comprehensive list of terms and definitions for greater clarification.

body of research, only procedural models are considered to be cognitive models. Figure 3 shows the relationship between physics-based and behavioral models with respect to the application areas of combat and OOTW. In general, physics-based models perform consistently in either combat or OOTW model applications with no differences in performance characteristics. For example, a model of an assault rifle maintains the integrity of the physical representation and physics of the weapon systems in either domain. Conversely, behavioral models may not perform consistently in either combat or OOTW model applications without noticeable differences in performance characteristics. For example, a model of human behavior for a combat model application cannot be federated with or integrated into another model of an assault rifle in the OOTW model application without altering the context and purpose of the physics model.

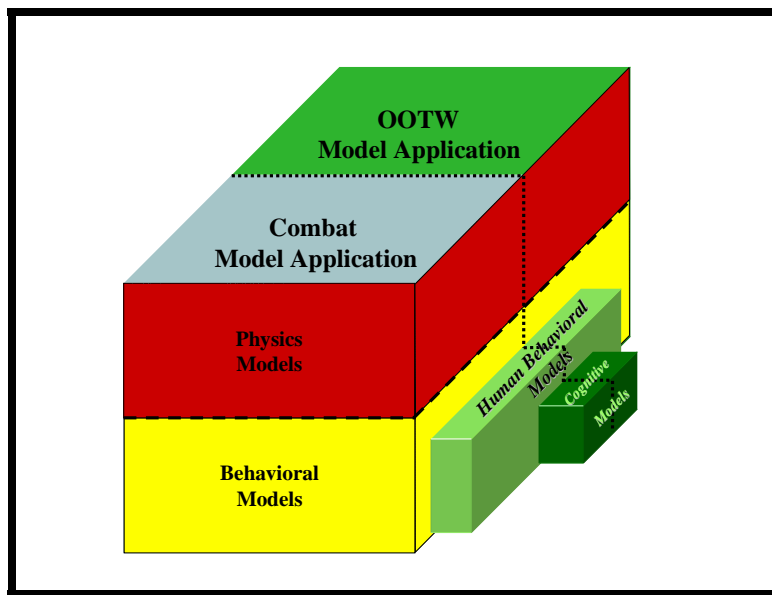


Figure 3. DoD Modeling and Simulation Landscape

A. VERIFICATION, VALIDATION AND ACCREDITATION

Verification, validation, and accreditation (VV&A) are important to ensure that models and simulations are ready for use.⁸ Verification and validation are generally conducted concurrently, with accreditation always being the final step in the process [DEPA 94]. Verification ensures model code and algorithms accurately represent the

⁸ For reference, key players involved in model VV&A for DoD use are provided in Appendix P.

real-world processes or objects modeled [DEPA 01a]. The Department of Defense Modeling and Simulation Office (DMSO) VV&A Technical Working Group (TWG) defines validation as “the process of determining the degree to which a model and its associated data are an accurate representation of the real world from the perspective of the intended uses of the model” [DEPA 01a]. Accreditation is an “official” seal of approval that the designated authority bestows on a model that confirms that the model has been properly verified, validated, and accredited for an intended purpose, application, and scenario.

Figure 4 depicts the iterative sequence of steps involved to VV&A DoD models and simulations. The process begins with identifying, defining, and scoping the problem. Next, an appropriate modeling and simulation method must be selected that is relevant to the purpose of the study and one that generates the right data for the decision-making process. Then a M&S plan is developed for building, verifying, validating, accrediting, and using the model. The model user must decide whether to use a legacy model as is, develop a new model, or federate multiple models together into a family of models.

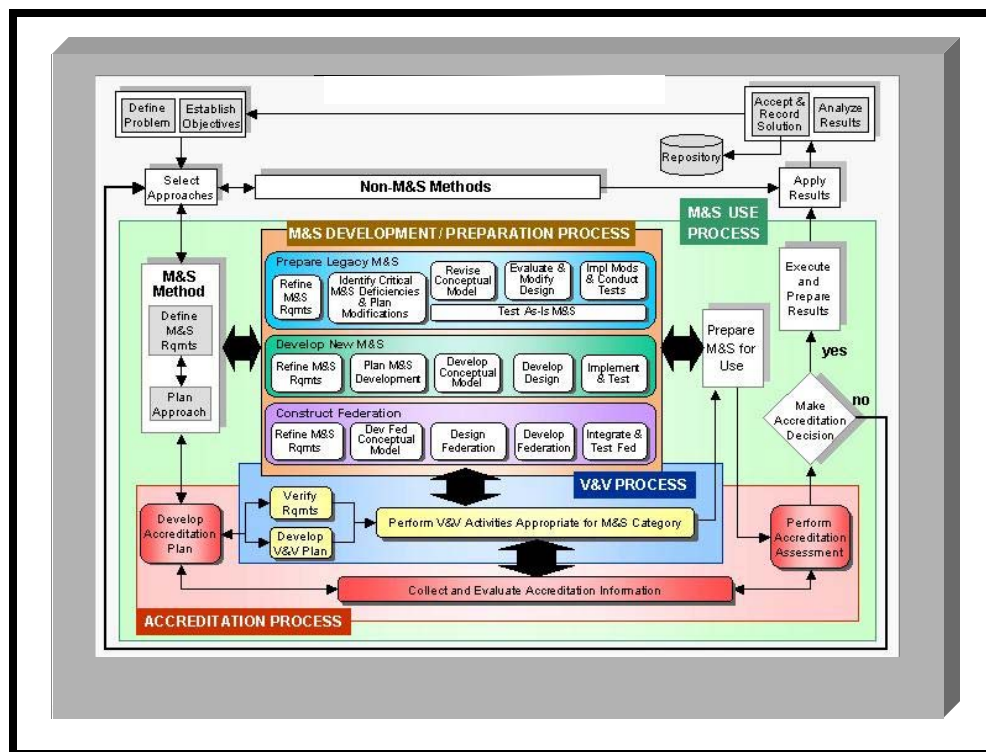


Figure 4. Modeling and Simulation Problem Solving Process From [DEPA 01e]

The verification and validation (V&V) process begins with a V&V plan that outlines V&V tasks given the type of model under construction. Specific requirements, tasks, and steps depend on the plan for building the model, the type of simulation the model is either integrated into or federated with, and the model's intended use [DEPA 01e]. Although V&V is required for both virtual and constructive simulations, it is understood, and common practice, to tailor V&V tasks to meet the unique needs and limitations of the model. DMSO's "Key Concepts of VV&A" list the following key V&V tasks [DEPA 01e]:

- Verify User Requirements;
- Develop a V&V Plan; and
- Perform the V&V Procedures Suitable for the Model's M&S Category: Validate Conceptual Model; Verify Model Design; Verify Model Implementation; and Validate Model Results.

Although DMSO's key tasks do not address requirements to validate data used to build and to test a model, the M&S community recognizes it is not possible to validate a model without test data to produce verifiable simulation results. For this reason, validating agents normally perform validation three times: referent, conceptual model, and model implementation [DEPA 01c].^{9 10}

Figure 5 illustrates where the validation steps fit into Birta and Özmiraç's model validation framework [BIRT 96]. Birta and Özmiraç do not explicitly address the roll of referent in their design of model development and model testing, however, referent is integrated into the diagram to show where it is created, validated, and used in model validation. Model implementation validation is the result of comparing simulation outcomes with real-world results under specific controlled conditions.

⁹ Validation agents are persons or organizations responsible for conducting validation of a model, simulation, or federation and supporting data [DEPA 01b].

¹⁰ Referent is the "codified body of knowledge about a thing being simulated" [HARM 98] [DEPA 01f].

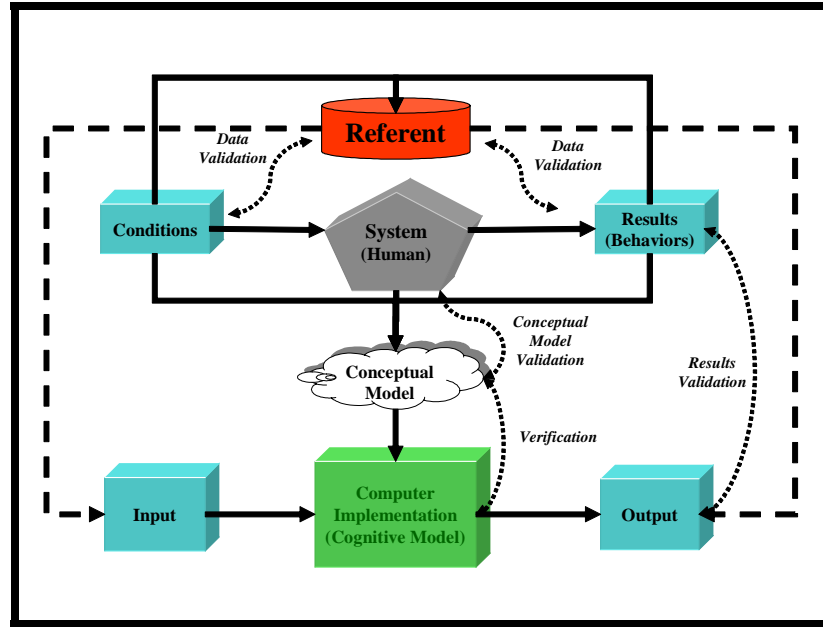


Figure 5. Birta and Özmırak Validation/Verification After [BIRT 96]¹¹

Data for the referent comes from many sources. One of these sources is validated models. Examples include models of specific aspects of human behavior, sociological phenomena, and the physiological processes underlying human behavior. Referent is also collected from validated simulations of human behavior (live, virtual, or constructive), empirical observations of actual operations, historical case studies, experimental data, and from SMEs [DEPA 01f]. Data also comes in various formats such as narrative, numerical, or tabular. Due to the source and nature of a referent required to build, validate, and operate models, numerous techniques exist for validating the referent. Table 1 lists five techniques used for validating referents. Validating agents may use combinations of these techniques to provide a more comprehensive validation. Table 1 identifies when it is most appropriate to use each technique from past M&S validation efforts.

¹¹ The original process proposed by Birta and Özmırak is modified in this document to reflect terms consistent with this research.

Table 1. Steps in Verification and Validation Process Where Comparison Techniques Best Apply After [DEPA 00b]¹²

Comparison Technique Class	Validation Process Step
SME Assessments	Conceptual model, data & face validation ¹³
Audits, Inspections & Walkthroughs	Conceptual model & data validation
Visual Comparisons	Data & face validation
Analytical Comparisons	Conceptual model & data validation
Formal Comparisons	Conceptual model, data & face validation

Table 2 presents a list of comparison validation technique limitations identified by DMSO. Limitations of comparison techniques illustrate an important aspect of validation plans and referents. Model requirements and specifications must be detailed and unambiguous. If they are not, the use of SMEs, auditors, and inspectors results in an unfocused validation effort. Comprehensive and explicit requirements and specifications scope the problem making model validation more manageable; however, they also focus the validation making it difficult to abstract the results and accredit the model for use in other domains.

Table 2. General Limitations of Different Comparison Techniques From [DEPA 00b]

Comparison Technique Class	Limitations
SME Assessments	<ul style="list-style-type: none"> SMEs should be available & properly prepared All information should be understandable to SMEs
Audits, Inspections & Walkthroughs	<ul style="list-style-type: none"> Teams should be properly composed, available, and prepared Sufficient information should be available for review sessions
Visual Comparisons	<ul style="list-style-type: none"> Information should lend itself to meaningful visualization Visualizations should be scaled correctly
Analytical Comparisons	<ul style="list-style-type: none"> Referents and requirements should be described in forms that permit comparison with model or simulation representations (e.g., UML)
Formal Comparisons	<ul style="list-style-type: none"> Information should take a formal, usually quantitative, form Uncertainties may need to be described but should absolutely be understood

Inconsistent or skewed data display can introduce a scaling effect when using *visualization comparison* techniques. This can distort validation results by exposing

¹² Knowledge base validation and other forms of complex data that the conceptual model may not represent fall under the term **data validation** [DEPA 00].

¹³ The original table labels face validation as results validation. Face validation is used in the dissertation to maintain consistence in terms.

SMEs to perception bias.¹⁴ Placing the data in proper perspective is often difficult and current technology limits the use of this technique. Therefore, validating agents normally use visualization comparison in conjunction with at least one other method validation technique. The degree of rigor and extensive resources required to use *analytical comparison* techniques make them less attractive than more informal techniques, however, they are excellent for validating conceptual models and knowledge bases due to their ability to investigate the composition and causality of models and simulations. Strictly defined specifications for extracting data used in *formal comparison* techniques make them the preferred means of verifying and validating a physics-based model's knowledge base, conceptual model, and results. However, the rigorous characteristics of this method limit the technique's applicability due to the time and money required to collect large amounts of data.

To assist the military M&S community with VV&A, the DoD has developed a series of instructions, regulations, and publications. Verification and validation procedures set forth in DoD and the three Services outline policies, assign responsibilities, prescribe general procedures, and provide a list of standard products required for accrediting a model. The documents do not provide a fixed set of procedures or a set of referent to validate models. The procedures follow the general phases outlined in Figure 6 and listed in detail in Appendix P. (Key Players in Verification, Validation and Accreditation).

In Figure 6, the clouds represent inputs into the system. *User objectives* help model developers characterize the requirements for the model. For example, an artillery battalion needs to have a cognitive model integrated into a new automated call for fire trainer (objective). The automated fire direction center (FDC) would need to interpret verbal calls for fire from forward observers (FO) (requirement). *Requirements* help developers filter through the *available referents* to identify the relevant referent(s) for use in developing algorithms and validating the final model. Developers do not use all referents during the development and initial testing of the model. Developers often place some referents aside for validation runs of the model. Examples of possible referents are

¹⁴ Performance bias is defined in Subsection II.F.2. Bias.

the ability to receive calls for fire, how to parse and evaluate call for fire messages, allocation of indirect fires, and time required to process a call for fire. *System information* provides insight to developers about the physical system or processes. Examples are weapon systems utilized, amount and type of ammunition available, ballistics of the ammunitions, and ammo/target pairings. Referent provides inputs for algorithms developed from the system characteristics to produce results. Validating agents compare these system results against the requirements using the validation referent. The final product is a set of documents that describe how well the codified model's results match the selected test referent.

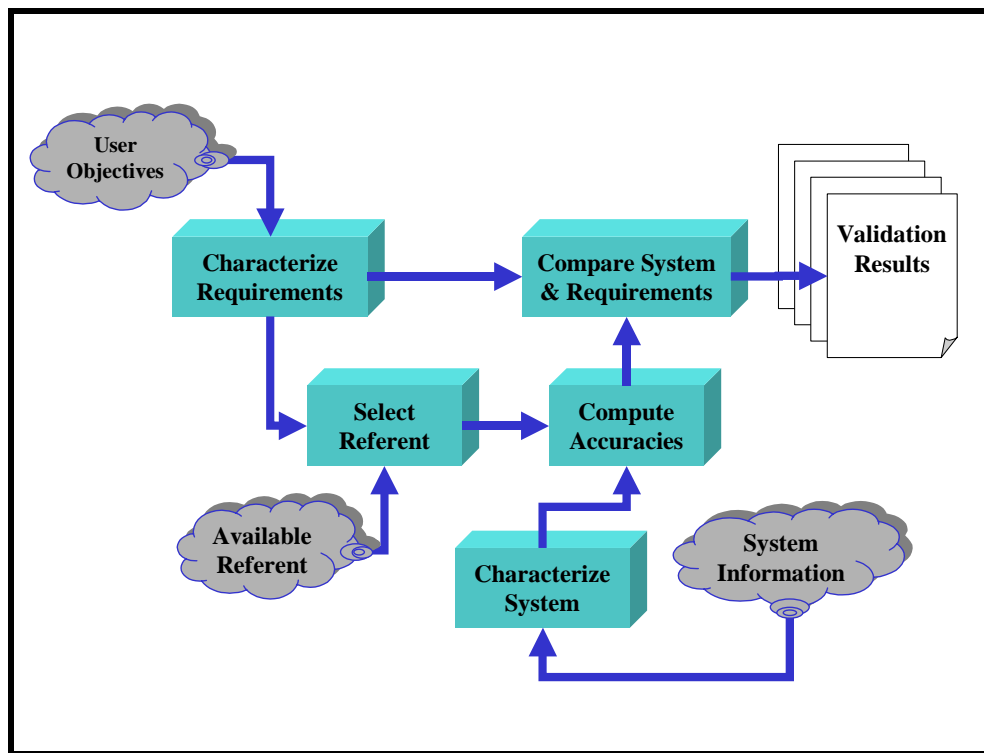


Figure 6. Essential Steps for Validating Models and Simulations From [DEPA 00b]

One of the most difficult phases of this process is the identification, collection, and selection of suitable referent to develop and validate the model. Table 3 presents four categories of information required for model validation and their associated sources. DMSO identifies SMEs as sources for three of the four categories. One of three is referents [DEPA 00b].

Table 3. Requirements and Sources of Validation Information From [DEPA 00b]

Validation Information Requirement	Information Sources
Requirements	SMEs, other user representatives, user documentation (e.g., concepts of operations)
Referents	SMEs, existing system documentation, experimental data, analysis and study reports
Model/Simulation	Conceptual model, design documentation, development team members
Comparison Techniques	Recommended Practices Guide (RPG), technical papers, SMEs

B. PSYCHOLOGY

The focus of psychology is the study of the representation and processing of information by complex organisms. It most often deals with species that process information in an intelligent manner. Intelligence implies the ability to obtain and process information in a manner that allows the organism to select behaviors with the best chance of “achieving the fundamental goals of survival and propagation” [WILS 99]. Previously, psychology focused on processing information amid sensory inputs and motor actions. Since psychologists consider humans “capable of the most complex and most domain-general forms of information processing,” most psychology research focuses on the nature of human intelligence and information processing [WILS 99].

One can see the diversity of psychology in its many fields or areas of interest. Behavioral psychology, cognitive psychology, cross-cultural psychology, and ecological psychology are four of these fields.¹⁵

Behavioral psychology deals with the study of overt responses to stimuli. Its focus is on overt responses to stimuli rather than on the mental processes. This focus failed to provide reasons for diversity in human behavior and neglected to account for elements such as “memory, attention, consciousness, thinking, and imagery” [SOLS 01]. In many cases, behavioral psychology rejected the theories of “mentalistic” [WILS 99].¹⁶ Previously, behaviorists attempted to operationally define these internal functions of the

¹⁵ Additional fields of psychology include: Clinical psychology, comparative psychology, developmental psychology, personality psychology, and social psychology.

¹⁶ Mentalistic refers to processes that are mental in origin (e.g. general knowledge, situational awareness, intent/goal, commitment, etc.) rather than physiological or physical [SHOH 93] [WOOL 95].

brain and roll them into a more general study of the mind [SOLS 01]. Although less popular than other areas of psychology, behavioral research continues today using many tools utilized by the natural sciences [WILS 99].

Cognitive psychology focuses on the scientific study of the human mind [WILS 99]. A cognitive psychologist studies how an individual or a group of individuals reasons through a problem. In doing so, the psychologist is concerned with perception, thought, and memory. Perception of knowledge deals with how an individual obtains information from the environment. Thought is concerned with how one solves problems and executes thoughts or relays thoughts to others. Memory involves the storage, retrieval, and processing of the information by the human brain. The domain of cognitive psychology is vast, covering as many as twelve principle areas: attention, cognitive neuroscience, consciousness, developmental psychology, human and artificial intelligence, imagery, language, memory, pattern recognition, perception, representation of knowledge, and thinking and concept formation [SOLS 01].¹⁷

Ecological psychology research deals with how an organism's behavior is based on its perception of the environment. This includes the shapes of objects, movement and change of objects, the organism's state and movement through the environment, and the organism's ability to influence the environment through effective actions. These

¹⁷ *Attention* is concerned with the ability to simulate input and/or process events stored in memory. *Cognitive neuroscience* is a study of how the mind-brain works at the level of the neuron. *Consciousness* deals with one's awareness of his/her internal or external conditions. Often deemed its own domain of psychology, some consider *developmental psychology* a subset of cognitive psychology. As discussed earlier, developmental psychology deals with how human behavior develops/changes over time. *Human and artificial intelligence* deals with recognizing and defining human intelligence so model developers can replicate it using a computer model. *Imagery* focuses on the mind's ability to take physical images to create a mental map from which the individual develops ideas and translates them into meaningful actions. The study of how humans learn and use *language* is often regarded as a subfield of developmental psychology. It concerns itself with the meaning of gestures and body posture as well as the written and spoken word. The field of *memory* research is involved with studying how the mind processes and stores events in short-term, working, and/or long-term memory. *Pattern recognition* is the study of how sensory inputs are grouped together to form recognizable patterns that are interpreted as a meaningful representation of information to be stored or retrieved from memory. *Perception* deals with "the detection and interpretation of sensory stimuli." [SOLS 01]. It attempts to determine how an individual takes sensory input and creates features and objects, categorizes and classifies these features and objects to develop a perception of the world. How information is represented, stored, and processed by the mind is the focus of *knowledge representation*. *Thinking and concept formation* is concerned with how thoughts and concepts are generated, confirmed, and modified.

perceptions differ for each organism. This is due to the ability of each organism to sense its environment and construct its own mental map of the world [WILS 99].

This is similar to how situational awareness or mental maps depict an individual's perception of the world. Situational awareness refers to a person's perception of the world based on sensory inputs, memories, and mental processing. One's situational awareness affects the actions one takes. Because of this, many cognitive models include a situational awareness module. Shattuck and Miller have been conducting research to address the effects of situational awareness on decision makers to determine measures of effectiveness for assessing the impact of systems designed to provide information for commanders to develop their situational understanding of the combat environment. [MILL 04]

Cross-cultural psychology "observes human behavior in contrasting cultures" where a culture is widely defined but routinely seen as pertaining to "patterns of behavior, symbols, and values" often transmitted over time [GALE 01]. This field of psychology asserts that the environment where an individual spends a great deal of time plays a dominant role in the behavioral patterns of an individual [MATS 99]. These patterns can influence everything from an individual's ability to extract information from symbols to how they perceive technology in general.

Cross-cultural distinctions can be large or small in scope. Psychologists consider global cultural characteristics based on environmental regions, religions, or systems of government as factors for cross-cultural studies; however, cultures can be even smaller. Examples of smaller cultural communities are branch of service (Infantry, Armor, Aviation, etc.) or unit type (light infantry, mechanized infantry, motorized infantry, or special operations). Psychologists may also use technology as a means of distinguishing cross-cultural characteristics. For example, categorizing behavior patterns based on three forms of technology exposure: Those who have never used computer technology, those who recently transitioned to the use of computer technology, and those raised with computer technology integrated into nearly every aspect of their daily lives. Prensky refers to these last two groups as digital immigrants and digital natives, respectively [PREN 01].

Understanding the varied fields of psychology allows us to investigate the impact of the various perspectives offered by the different fields within psychology on HBR models. The procedural aspect of behavioral psychology can be seen in many of the rule-based implementations of modern HBR models where models abstract responses based on stimuli with limited consideration for the thought process behind those decisions. One can also see this abstraction in the use of face validation techniques to validate the overt results of HBR models.

The cognitive psychologist Wilhelm Wundt heavily used *introspection* in the 1880s and 1890s. His method required trained observers to analyze “their own thought processes as they performed various cognitive tasks” [WILS 99]. This self-analysis often lead to biased results, skewed towards how observers were prone to hypothesize. Because of its inconsistencies and apparent lack of objectiveness, many psychologists viewed introspection as “unscientific.” Behavioral psychologists were some of the first psychologists to rebuke introspection techniques as a valid means of collecting data [RUSS 95]. Since the 1930s, its use in the field of psychology for collecting information has been limited [WILS 99].

Today, research personnel use a modified version of introspection, cognitive task analysis, to collect information about a specific domain. However, instead of using observers trained in the field of psychology as sources of information, SMEs are the source of information and psychologists collect the data. As with introspection, biases may impact data collected, however, this bias is based on preferred techniques of SMEs, SMEs developing cognitive maps that differ from the facts presented, and the training effect of SMEs reviewing numerous tasks and scenarios.

HBR models, as used by psychologists, are tools to represent observations and assumptions of how the mind works. Psychologists use HBR models to explain a specific theory, further research in cognitive psychology, and study complex concepts of storage, retrieval, and processing of memories. HBR models help to develop hypotheses and make behavioral predictions. One of the most famous and simplistic cognitive models is Waugh and Norman’s model of human memory (Figure 7).

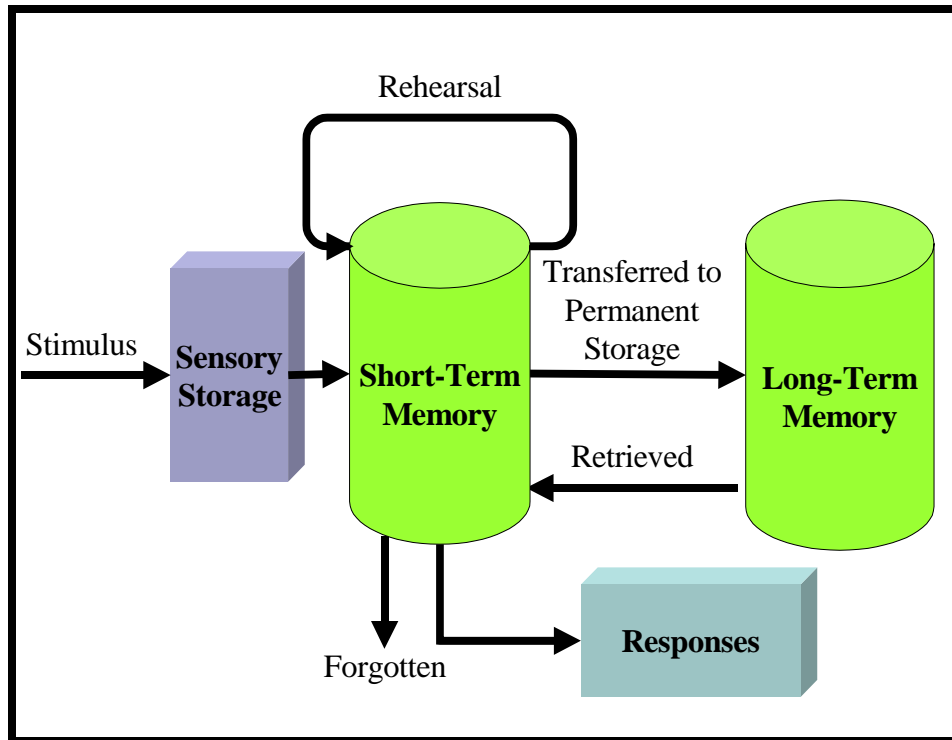


Figure 7. Waugh and Norman's Model of Human Memory From [SOLS 01]

Many cognitive architectures use variations of this human memory model to represent the storage and retrieval of facts. Understanding this theory may lead to better techniques for validating HBR models as we identify the types of information stored in each section of the model, when a segment of memory is accessed to make decisions, and when memories are lost or are inaccessible. Other constraints may limit the search for an optimal decision where the decision maker abandons or bypasses more formal thought processes to quickly select a plausible solution.

One can see a commonality between ecological psychology and the manner in which military decision makers address situations based on a leader's prior assignments. Lessons learned and techniques used in previous assignments may lead decision makers to recognize certain enemy behavior patterns and select a behavior to address the perceived situation. Examining research techniques used in the field of ecological psychology may provide insight into new methods of identifying ways to represent situational awareness in HBR models, the fusion of information, and presentation of the common operating picture in combat simulations.

Another issue is SME bias based on ecological and cross-cultural influences. This is present in both the development of a HBR model and the collection of referents when using SMEs. This bias discounts possible options based on the way people were raised and trained to think, the region of the world an individual was reared, and other cultural influences which affect an individual's performance. An example of such influence is the value people place on a human life. Cultures who place a relatively higher value on a single life may not consider the option of using suicide bomber(s). On the other hand, a culture which values the well-being of the majority over a single life, may see suicide bombers as a viable option to its current dilemma. The reasoning processes of individuals in each culture may lead to seemingly dissimilar behaviors.

C. COGNITIVE MODELS

As with model taxonomies, cognitive models can be described at three different levels: representations, architectures, and implementations.

1. Representations and Architectures

As stated earlier, cognitive models deal with the human decision-making process. Cognitive model representations provide a means of describing different methodologies for representing codified cognitive functionality. Codified cognitive modeling has been the focus of two major communities over the past fifty years, artificial intelligence and artificial life.

The artificial intelligence (AI) community has numerous goals but in general, the focus has been on comprehending intelligent computerized entities [RUSS 95]. The techniques used by the AI community generally involve a top down approach requiring an attempt to codify all relevant behavioral details [RALS 00]. These techniques use inductive and deductive reasoning to identify and codify entities to display rational behavior (correct actions) [RUSS 95]. The emergent field of artificial life (AL) attempts to model the behavior of biological systems [FREE 99]. The AL community uses a bottom-up approach to identify and codify characteristics in computer entities allowing entities to evolve and emerge to perform intelligent actions. The focus of AL is emergent behaviors of entities as they attempt to survive in complex environments [RALS 00].

The two communities have developed numerous techniques for implementing their approaches. Some of these techniques fuse the boundaries between AI and AL, such

as multi-agent systems, while others are contained primarily in one domain. Examples of cognitive model representations are Agent-Based, Bayesian-Network, Multi-Agent System, Neural-Networks, and Rule-Based.

Agent-Based representations demonstrate intelligence through codified objects that perceive characteristics of the environment and act on those perceptions [RUSS 95]. There are several types of agent-based cognitive architectures. Two of these are reactive and rational agents.¹⁸ A reactive agent bases its actions solely on the last set of sensory inputs. Often the approach uses a simple condition-action rule (e.g., this is my perceived state of world; I choose this action). A rational agent uses sensors to perceive its environment and performs actions on the environment using effectors. Rational agents maintain a state of situational awareness based on their past knowledge of the world and current sensory inputs [RUSS 95].

The *Multi-Agent System* (MAS) is a relatively new representation for replicating behaviors based on the Complex Adaptive System (CAS) theory. Developed in the late 1970s, MAS is a system with autonomous or semi-autonomous software agents that produce adaptive and emergent behaviors.¹⁹ The model uses a bottom-up approach where software agents have independent micro-decisions that generate group level macro-behaviors. A MAS can use any form of agent-based software technology (reactive, rational, goal-based, utility-based, etc.) with the agents characterized as possessing intentions that influence their actions. Multi-agent systems are used in large domains where non-linearity is present [HOLL 95]. The MAS, limited only by the physics constraints of the simulation boundaries, uses an indirect approach to search the large domain for viable results. Another feature of MAS is its ability to allow agents to evolve to create new agents which, in general, are more optimized to survive/thrive in the simulated environment [FERB 99]. If coded with a *brain lid*, one can interrogate agents

¹⁸ Russell describes agents as three types: reflex agents or reactive agents, goal-based agents that attempt to achieve a specified goal, or utility-based agents that attempt to achieve the best possible state from their point of view [RUSS 95].

¹⁹ Adaptive behavior is the process of fitting oneself to the environment. A MAS generates emergent behavior at a higher cognitive level based on the behaviors and interactions of agents at a lower level. Schelling describes this as micro decisions leading to macro behaviors [SCHE 78].

for the reasoning behind their actions as well as view their overt behaviors [LEWI 02].²⁰ Examples of MAS are the Irreducible Semi-Autonomous Adaptive Combat (ISAAC), Pythagoras, Socrates, Enhanced ISAAC Neural Simulation Toolkit (EINSTEIN) and Map Awareness Non-uniform Automata (MANA) [ILAC 97] [PROJ 02].

Cognitive model *architecture* is the framework for establishing how the components of the cognitive model relate to each other. Cognitive model architectures use one or more cognitive model representations to structure the schema behind a specific cognitive model. An architecture is not a functioning model implementation, but the design for an implementation. Examples of cognitive model architectures are the Adaptive Control of Thought (ACT-R), COGNition as a NETwork of Tasks (COGNET), Connector-Based Multi-Agent System (CMAS), Executive-Process Interaction Control (EPIC), and State, Operator And Result (Soar). Table 4 indicates some of the means by which these architectures can provide information to explain their actions. Each architecture can demonstrate its overt behaviors, but most are limited to their ability to provide information about the specifics behind the cognitive processes they used for their behavior selection.

²⁰ Programmers code a *brain lid* into an agent to allow inspection of the agent to determine its situational awareness and decision processes leading to a specific action [RODD 00].

Table 4. Model Architecture Action Information Sources After [PEW 98] [OSBO 02]

Model	Information for Action Explanation
ACT-R	<ul style="list-style-type: none"> ▪ Overt Behaviors ▪ Encoded Knowledge ▪ Encoded Rules ▪ Decision Stack ▪ Declarative knowledge used ▪ Changes in working Memory ▪ Final Parameters ▪ New Rule & Productions ▪ New Declarative Memory
CMAS	<ul style="list-style-type: none"> ▪ Overt Behaviors (Actions) ▪ Goals ▪ Tickets (Possible Actions to achieve a specific goal) ▪ Outer Environment (State of the model) ▪ Inner Environment (An agents Situational Awareness) ▪ Entity State ▪ Connectors (Possible entity interactions)
COGNET	<ul style="list-style-type: none"> ▪ Overt Behaviors ▪ Conditions/ Rules ▪ Blackboard (Situational Awareness)
EPIC	<ul style="list-style-type: none"> ▪ Overt Behaviors ▪ Encoded Knowledge ▪ Encoded Rules
Soar	<ul style="list-style-type: none"> ▪ Overt Behaviors ▪ Encoded Knowledge ▪ Decision Stack ▪ Knowledge Stack

2. Implementations

A cognitive model implementation takes a generic cognitive model architecture with its supporting cognitive model representation(s) and provides code and data for each component. An implementation is a functional representation of the architecture.

Ilachinski created the *Irreducible Semi-Autonomous Adaptive Combat* (ISAAC) model in 1997 for the U.S. Marine Corps to investigate the utility of agent-based systems. One of the goals of ISAAC is to show that land combat can be modeled using a CAS. As an implementation of AL, ISAAC introduces dynamic emergent behavior in an attempt to overcome shortcomings of Lanchester-type combat models. [ILAC 97] As an AL implementation, ISAAC exhibits the effects of a model with no central control; the interaction between autonomous or semi-autonomous entities often produces unpredictable outcomes. The model attempts to fill some of the perceived gaps between

the current needs of the M&S community and the shortcomings of previous HBR implementation to represent dynamical human behaviors.

The model uses agents with four properties to generate believable behavior:

- Embedded "doctrine" is a default set of local-rules used to specify how an agent is to act in a generic environment
- A "mission" is a goal directing behavior
- "Situational awareness" results from sensors generating an agent's internal perception of the environment
- Behaviors and/or rules are altered through an internal adaptive mechanism [ILAC 97]

The system can run in an evolutionary mode utilizing a genetic algorithm to increase an agent's ability to survive.²¹ Using the evolutionary mode of operation, ISAAC has shown an impressive catalog of emergent behaviors. This list includes the ability to perform a frontal attack, local clustering, penetration, retreat, containment, flanking maneuvers, and encirclement of the enemy [ILAC 97].

The *Map Awareness Non-uniform Automata* (MANA) model is another model in the Marine Corps Combat Development Command's (MCCDC) Project Albert. Project Albert is the Marine Corps' research effort to assess the general applicability of the use of CAS to study land warfare. Other HBR models in Project Albert include Pythagoras, Socrates, and ISAAC [PROJ 01] [PROJ 02].

The Defence Technology Agency of New Zealand developed MANA to conduct research into implications of chaos and complexity theory for combat and other military operational modeling.²² MANA is an agent-based representation developed based on Enhanced ISAAC Neural Simulation Toolkit (EINSTEIN) and its precursor ISAAC.

As with other agent-based models (ABM), MANA consists of entities controlled by decision-making algorithms. The model's developers further classify MANA as a

²¹ A genetic algorithm searches the collection of individual agents to find the agent that maximize the fitness function and then uses the agent(s) to produce new agents. The fitness function takes the agent as an input and delivers a numerical output based on the agent's internal state and resulting performance function. A fitness function can be derived from anything configurable as an optimization problem. [RUSS 95]

²² The following description of MANA is drawn directly from the *MANA, Map Aware Non-uniform Automata, Version 3.0, Users Manual (Draft)* [GALL 03]

CAS. MANA's entities represent military units which make decisions based on a "memory map" which provides individuals or entities with goals to guide them about the battlefield.

Some of the aspects that allow MANA to be designated as a CAS are:

- MANA has the ability to exhibit "global" behavior, materialized based on local interactions;
- MANA uses feedback to update agents regarding changes to the environment;
- MANA cannot be analyzed by decomposing it into simple independent parts; and
- Similar to human behavior, agents "adapt" to their local environment and interact with each other in a non-linear manner.

MANA has the ability to incorporate several additional features which ISAAC did not have when MANA was initially developed. These include:

- Shared memory of enemy contacts provides agents with enhanced situational awareness. MANA uses two mechanisms to provide situational awareness, "squad map" and "inorganic map". The "squad map" maintains group contact data. The "inorganic map" stores contacts based on communications from other units.
- Communications exists between units in order to pass contact information. The model can alter information accuracy based on the influence of unit activities and environmental conditions on communications.
- *Terrain Maps* contain features such as roads which increase agent speed and undergrowth which agents can use for concealment.
- The use of waypoints for routes provide intermediate goals to facilitate coordination of units and achievement of an ultimate goal.
- Agent personalities can be event-driven. Events (e.g., making enemy contact, being shot at, engaging others, reaching a waypoint, etc.) can activate a special personality trait, present for a limited amount of time or until modified by another event. Personality changes can be set individually or for an entire unit.

MANA divides its parameters into four categories: personality weightings, move constraints, basic capabilities, and movement characteristics. *Personality weightings*, determine an automaton's propensity to move towards friendly or enemy units, towards its waypoint, towards easy terrain, and towards a final goal point. Next, *move constraints* act as conditional modifiers. An example of a modifier is the "Combat" parameter, which

determines the minimum local numerical advantage a group of agents needs before approaching the enemy. *Basic capabilities* describes an agent based on its use of weapons, its use of sensors, its movement speed, and its tendencies for interaction with other agents. Finally, *movement characteristics* of the agents, include the effects of terrain on agent speed, the degree of random agent movement, and agent's desire to avoid obstacles. [GALL 2003]

D. HUMAN BEHAVIOR REPRESENTATION

Human behavior representations (HBR) model human behavior at one of four levels: combined organizations, organizations, individuals, or components of individual performance. They may represent one or more cognitive functions such as perception, inference, planning, or control. HBRs can also portray the effects of behavior modifiers: stress, injury, fatigue, discomfort, motivation, and emotion. They often have human performance restrictions such as decision latencies or bandwidth allocated for sensing [DEPA 01f].

Within DoD M&S, HBRs are referred to as one of the following:

- Automated FORces (AFOR),
- Command FORces (CFOR),
- Computer Generated Forces (CGF),
- Semi-Automated Forces (SAF and SAFOR), or
- Synthetic forces [DEPA 01f].

1. Human Behavior Representation Verification and Validation Procedures

Although the purpose and implementation of physics-based and HBR models are fundamentally different, the V&V processes are the same. The validating agent must evaluate the capabilities of the physics-based and HBR model at four discrete phases. Figure 8 is a graphical depiction of the four phases of model development and the high-level validation tasks that DMSO defines as necessary for a validation agent to perform a comprehensive validation of a an HBR model: (1) conceptual model design; (2) contents of the knowledge base; (3) implementation of the model and its knowledge base; and (4) integration of the model into the simulation. The degree to which the validating agent can validate a model in each phase is dependent on the model representation. Representations

such as neural networks can only undergo face validation due to the complexity of the underlying model, which validating agents often treat as a “black box” [DEPA 01f]. Within the four phases, HBR VV&A requires the completion of several high level tasks is essential:

- (a) Collecting a complete a set of requirements and acceptability criteria;
- (b) Identify referents for in assessing the HBR’s validity;
- (c) Validate conceptual model against the requirements using the referents;
- (d) Analyze conceptual model to identify areas of high complexity to focus model implementation validation efforts;
- (e) Validate knowledge base against requirements using referents;
- (f) Analyze knowledge base to identify areas of high complexity to focus model implementation validation efforts; and
- (g) Validate integrated HBR implementation against requirements using referent and concentrating on key areas identified during the conceptual model and knowledge base analysis [DEPA 01f].

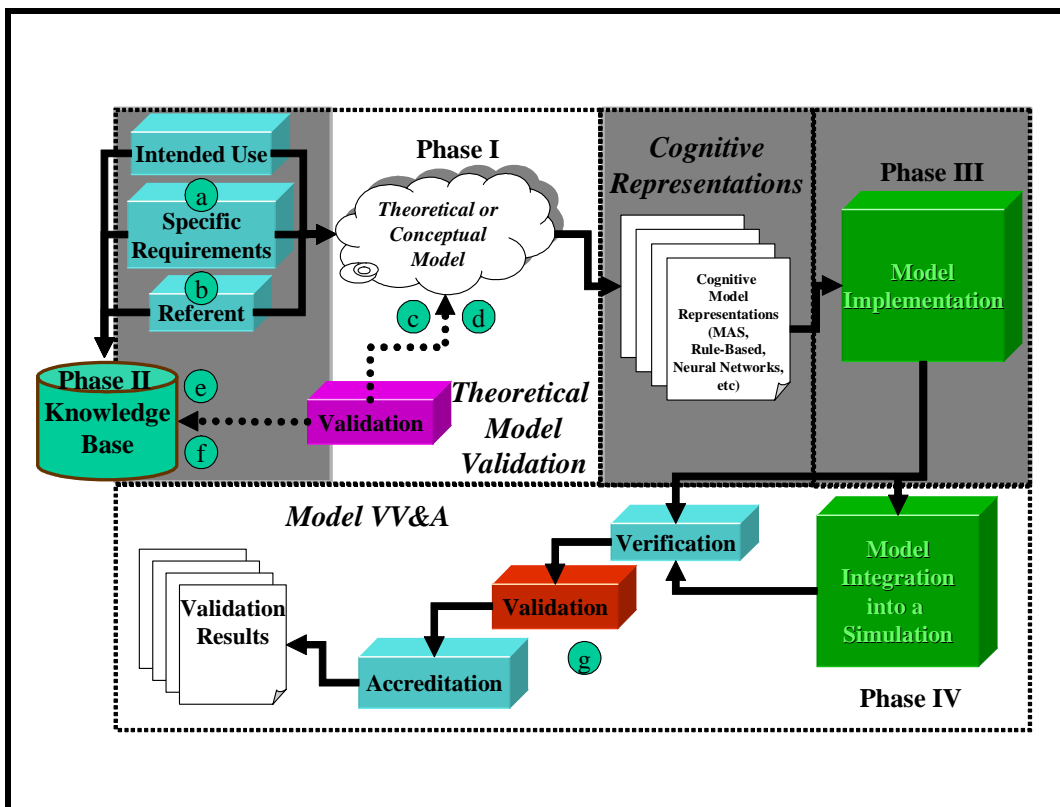


Figure 8. Verification, Validation, and Accreditation Tasks for a Human Behavior Representation Model After [23]

²³ See [DEPA 00b] [DEPA 01e] [DEPA 01f]

Prior to use, the model must be validated. For physics-based models, this normally includes completing a proof and conducting a numerical validation of the model. For HBR models, SMEs normally perform the less quantifiable and more qualitative method of face validation on the conceptual model to determine if the model has any major theoretical faults based on the current understanding of the human thought process. This research assumes the cognitive architecture is valid, and focuses on face validation of the coded implementation of the HBR model.

2. Referent Categories

There are numerous ways to categorize referents. DMSO's "Key Concepts of VV&A" section of its Recommended Practices Guide (RPG) describes six categories of correspondence useful for determining referent for HBR: computational correspondence, domain correspondence, physical correspondence, physiological correspondence, psychological correspondence, and sociological correspondence [DEPA 01f].²⁴

Viewing the human mind as a machine made of an immense assortment of computational devices, *computational correspondence* addresses the ability of the human nervous system to take inputs, process the inputs, store information, retrieve stored information, make decisions, and produce outputs. Cognitive psychologists commonly accept that the brain performs these functions, however the physical specifics of how the brain performs these tasks is not well understood. However, psychological studies have identified bandwidth and storage limitations of the human brain for specific tasks. Validating agents have used this referent in conjunction with theories of brain computational performance to conduct limited validations of cognitive models [DEPA 01f].

Domain correspondence addresses the use of SMEs to examine the knowledge base and outcomes of human behavior in their specific area of interest. The data collected is normally qualitative and leads to referent viable for face validation. Researchers often equate this form of validation to a Turing Test [DEPA 01f]. This referent is generally gathered from the research of behavioral psychology.

²⁴ Correspondence is the agreement of a model to different levels of abstraction.

Comparing the results of physics-based models against human physical constraints is known as *physical correspondence* [DEPA 01f]. This referent is normally limited to the more obvious physical constraints of the human body (e.g. how fast a human can run, how much a human can carry, etc.).

Physiological correspondence resembles data used to validate physics models. It uses information from neurologists, neurosurgeons, or physiologists to determine if a model's components react similar to the portion of the brain they simulate. This form of validation has become more viable over the last two decades due to advances in understanding of the physiology of the human nervous system. Physiological correspondence is an immature area of study but it has demonstrated use in validating neural networks [DEPA 01f].

The SME for *psychological correspondence* is the psychology professional. Similar to SMEs and domain correspondence, psychologists provide qualitative analysis of the real-world behavior and model results to determine if the model exhibits human-like behaviors. One can mine data to support psychological correspondence from the numerous volumes of experimental data on human performance in varying real-world scenarios [DEPA 01f].

Validating a model using psychological correspondence has potential issues with the qualitative nature of the referent and unintentional bias of the psychological experts, similar to that identified in introspection. However, psychological correspondence testing has the potential for greater credibility as the M&S and Psychology communities codify and validate more models of emotional phenomena. These validated models may provide baseline data and reduce the need for an exhaustive search of psychological problem space to identify appropriate referent. This shows most promise for models that incorporate aspects of stress and emotion [DEPA 01f].

For cognitive models of group behavior, *sociological correspondence* provides data on the interactions between groups and individuals. It includes groups operating under a unified organizational structure and unordered groups (crowds, mobs, etc.). An extensive body of knowledge exists from simulated and real-world situations from which one can acquire referent on sociological correspondence. The body of knowledge

includes interactions between groups, between groups and individuals, and between individuals in groups. Sociological correspondence also has the luxury of well-established experimental protocols of sociological experiments to develop validation tests [DEPA 01f]. This form of correspondence is closely related to cross-cultural psychology.

3. Face Validation

To date, the most common means of validating cognitive models has been through face validation using SMEs [DEPA 00b]. Often this technique uses a SME to exercise the HBR in a scenario where the SME manipulates the model through the simulation space by issuing orders or varying the stimulants, observing the resulting behavior, and determining whether the behavior meets a user's requirements for realism. SMEs often use personal opinions or qualitative referent provided by validating agents for face validation of HBR models [DEPA 01f].

Harmon and Metz propose new criteria for the validation of HBRs. They believe a strict level of validation for HBR models is idealistic. Harmon feels establishing a set of validation levels for the validation of an HBR would provide the M&S Community with a more meaningful and attainable validation process for HBR models [HARM 03]. Goerger, who concurs that a single validation standard for all HBR models is impractical, proposes a sliding scale of validation to indicate the flexibility of an HBR model [GOER 02] [GOER 03].

4. Subject Matter Experts

The Defense Modeling and Simulation Office VV&A TWG provides a list of general attributes individuals should demonstrate if they are to be used as SMEs [DEPA 00a]. These traits include independence, recognized competence, trust, good judgment, and perspective [DEPA 00a].²⁵ Pace and Sheehan feel these five traits fall short of providing standardization for SME certification. They propose more ridged guidelines for

²⁵ Independence suggests that a SME is impartial and can provide an "honest and probing assessment". A SME is one with the level of experience and knowledge of the subject matter and process to perform the task(s) the validating agent is asking him to execute. Trust is the "confidence that an SME has no hidden agenda detrimental to the simulation development." Good judgment indicates a SME can judge when he (or his team) has sufficiently examined the model to provide a proper assessment of its capabilities and limitations. Perspective is a SME's ability to maintain focus on the objective and limitations of the validation effort [DEPA 00a].

SME certification similar to those used by the judiciary system to classify individuals as expert witnesses. Such standards of excellence could help to ensure the legitimacy of a SME pool [Pace 02].

As described earlier, model developers use SMEs throughout the VV&A process to perform tasks such as collecting data, validating the knowledge base, validating the theoretical model, and validating the model implementation. The use of SMEs to perform face validation is analogous to the use of introspection. Despite the limited use of introspection in psychology, validating agents still use “behavior visualization techniques (which are similar to introspection, because these techniques) can greatly help SMEs examine simulation results, particularly for simulations with which they (the SMEs) can interact.” [DEPA 00b]

5. Issues

Although preferred, formal validation is not always attainable. “Current state-of-the-art proof of correctness techniques are simply not capable of being applied to even a reasonably complex simulation model. However, formal techniques serve as the foundation for other V&V techniques.” [BALC 97] Because multiple V&V agencies with non-standard criteria or non-uniform referent perform validation, validating agents inconsistently apply the validation process [DEPA 01f]. This often leads to an invalid comparison of cognitive models due to the non-uniform means of validation and inconsistent validation efforts.

The high-level V&V tasks and issues with referents lead to other innate difficulties in validating human behavior models. DMSO has identified four factors, making validation of HBR models difficult. First is the very large set of possible actions for the simplest human behaviors. This makes it difficult to ensure complete consideration of all viable solutions. Second is the general non-linear characteristic of the constrained space of consideration. The non-linearity of the space prevents a simple causal relationship to be drawn between situational parameters and resulting actions. Next is the tendency of behavioral model developers to use stochastic algorithms in HBR models to demonstrate unpredictability. This ‘unpredictable’, unless it can be made deterministic, typically makes repeatable runs of the model impossible. Therefore, the

model becomes difficult, often impossible, to validate. DMSO's fourth hindrance to validation is the chaotic behavior exhibited by HBR model implementations that are sensitive to initial and boundary conditions. Models with such sensitivity issues are limited to the breadth of their validation to the subset of scenarios where they exhibit stable behavior [DEPA 01f].

E. VALIDATION EFFORTS OF HUMAN BEHAVIOR MODELS

Over the years, the M&S and psychology communities have developed numerous HBRs for a variety of purposes. The National Research Council conducted a study in 1988 to review the state of HBR and organizational modeling. One of the products of the study is a survey of validation efforts for many of the HBRs in existence or under development at the time. Table 5 summarizes and compares the different HBR validation approaches discussed in the study [PEW 98].

Table 5 includes the domain for which each cognitive model was developed, the types of correspondence used for validation, and the sources of referents. Correspondence categories were limited to either domain, physiological, or psychological based on the techniques employed by validating agents at the time of the report. As stated earlier, domain and psychological correspondence gather their referents from SMEs. The use of SME-derived referents makes these two forms of validation subject to bias, frequently limited to qualitative data, and routinely resulting in face validation of the model. Models validated using more than one category of correspondence often focus on domain and psychological correspondence, which are typically limited to face validation of overt behaviors.

Table 5 illustrates the difficulties in comparing models based on their validation efforts since not all models are validated using the same techniques or correspondence. It also expresses the need for developing standardized procedures for the validation of HBR models to ensure model users provided more than a cursory review of the model prior to their use in a simulation. Finally, the table indicates the difficulty in collecting referents for each category of correspondence for use in developing and validating HBR models for different domains. While not the easiest data to collect, human performance data is definitely an area in which the DoD has focused a majority of its referent collection resources.

Table 5. Comparison of the Validation of Different HBRs From [PEW 98]

Cognitive Model	Domain Types	Correspondences			Validating Data Sources
		Domain	Psychological	Physiological	
ACT-R	submarine TAO & Aegis radar operators	X	X		• human behavior data
COGNET	anti-submarine warfare	X			• human behavior data
EPIC	computer interaction tasks	X	X		• human behavior data
HOS			X		• validated theory
Micro SAINT	helicopter crew, ground vehicle crews, C2 message, tank maintenance & harbor entry operations	X			• human behavior data
MIDAS	757 flight crew	X			• human behavior data
Neural Networks			X	X	• validated theory • human behavior data
OMAR			X		• validated theory • human interaction
SAMPLE			X		• validated theory
Soar	air traffic control, test director, automobile driver, job shop scheduling	X	X		• validated theory • human interaction • human behavior data
ModSAF	ground warfare	X			• human interaction
CCTT SAF	ground warfare	X			• human interaction
MCSF	small unit operations	X			• human behavior data • human interaction
SUTT CCH	small unit operations	X	X		• human behavior data • human interaction
IFOR (see Soar)	fixed & rotary wing air operations	X	X		• validated theory • human interaction • human behavior data

All validation techniques have limitations. The cognitive models listed in Table 5 indicate there are two significant limitations of HBR correspondence used for validation. First is the unrealistic requirement of domain correspondence to search very large and nonlinear behavior spaces. For example, identifying and codifying every factor influencing a soldier's decision on a dismounted route through the woods, swamp, jungle, desert, arctic, or urban terrain includes elements of mission, enemy, terrain, time, troops, weather, equipment, etc. Second concerns testing for psychological and physiological correspondences. These two forms of correspondence usually require the use of

extensively validated models of psychological and physiological phenomena to produce referent [DEPA 01d]. In essence, one must find results from other valid HBR models or build and validate another HBR model to provide referents for validation of a new model. This dependence on other models makes validation using psychological and physiological correspondences tenuous at best.

F. HUMAN PERFORMANCE EVALUATION

Supervisors evaluate personnel for two reasons. First is to determine who is due just rewards and promotions. Second is to determine what additional training is needed to help develop individuals and teams [TZIN 00]. This process is complex and fraught with potential issues which human resource personnel have established techniques to help resolve. To address some of these issues and techniques, the remainder of this subsection covers the fundamental elements of human performance evaluation, the common problem of evaluator bias, and some of the possible techniques shown to mitigate bias.

1. Procedural Versus Declarative Knowledge

Knowledge normally used to provide input to human performance evaluation is categorized as either declarative or procedural. Declarative knowledge is facts -- the “what”. Examples of declarative knowledge are an M16A2 is a semiautomatic rifle used by the US Army, an M16A2 semiautomatic rifle uses a 5.56mm round, and an M16A2 can fire a using 3-round burst or single shot modes. Procedural knowledge involves comprehension of the process -- the “how”. For example, before firing an M16A2, one must load the weapon by inserting a magazine containing one or more rounds of ammunition, allow the bolt to slide forward to chamber a round, and move the shot selection switch from safe to single shot or burst mode.

Procedural knowledge is declarative knowledge interpreted within the context of situational understanding. Without declarative knowledge, procedural knowledge has no foundation. Without procedural knowledge, declarative knowledge is limited to the statement of facts. This difference allows one to look at an incident in two ways. Declarative knowledge allows you to collect the facts of what happened, while procedural knowledge allows you to determine why it happened. This is illustrated by comparing overt behaviors with cognitive processes. Overt behaviors are described as declarative knowledge, while cognitive processes allow the user to understand why a

particular behavior was selected. A combination of the two categories permits supervisors to provide a more complete assessment of personnel by demonstrating if the sum of the facts is equal to the whole. This explains why assessment requires context and not just analysis of the raw facts.

2. Bias

As defined by Webster's Dictionary, bias is "systematic error introduced into sampling or testing by selecting or encouraging one outcome or answer over others." [MERR 03] Bias often occurs in the assessment of human performance. Research literature describes at least five types of bias applicable to SMEs: judgmental, decision, heuristic, informational, and normative.²⁶ One can further classify judgmental and decision bias into at least twenty subcategories: anchoring, adjusting, association, availability, base rate neglect, belief, certainty effect, central tendency, confirmation, conjunction, conservatism, contrast, framing, halo, hindsight, illusory correlation, insensitivity to the prior probability of outcomes, leniency/severity, overconfidence, regression to the mean, representativeness, response bias, sunk costs, and the Law of Small Numbers [TVER 71] [TVER 74] [KAHN 82] [COHE 93] [BARN 93] [PERR 93] [CASA 98] [STEI 98] [GILO 02].²⁷

Pace and Sheehan categorize bias associated with the use of SMEs into three dimensions: perspective, performance, and perception [PACE 02]. *Perspective* addresses a SME's ability to maintain focus on the intended purpose of the model. A SME may lose focus as he allows his real-world experiences to cloud his view on what the model should have the capability of doing. *Performance* deals with the SME's ability to execute the validation process. This ability may be hindered by demands on the SME's time, the availability of data, the SME's ability or desire to comply with specified validation procedures, or the ability of the expert to understand the simulation. Finally, *perception* addresses the bias an expert brings to the process based on his education, training,

²⁶ The Glossary provides definitions for each bias category.

²⁷ This work only defines those subcategories specifically addressed in this dissertation: anchoring, contrast, confirmation, and the Law of Small Numbers. The remaining subcategories are listed to provide an indication of the vast number of bias which might effect evaluation results.

real-world experiences, exposure to simulations, and organizational loyalties. These factors may unduly focus a SME's attention on certain aspects of a model's performance [PACE 02].

Three subcategories of perception bias, which this research addresses, are anchoring, contrast, and confirmation. *Anchoring bias* emerges when an individual embraces an initial hypothesis and maintains this view regardless of incoming facts. This results in overemphasis on the hypothesis and an inappropriately minimal shift from the initial viewpoint [TVER 74] [KAHN 82] [COHE 93] [DUFF 93] [PERR 03] [STEI 98]. *Contrast bias* materializes when one seeks information to contradict an original hypothesis, ignoring or undervaluing evidence in support of the hypothesis [TVER 74] [KAHN 82] [PERR 03]. *Confirmation bias* is demonstrated when an individual overvalues select pieces of information relative to consistent evidence indicating an alternate conclusion [COHE 93] [DUFF 93] [PERR 03] [STEI 98].

Subject matter experts show bias on many levels. One characteristic of a SME is his ability to quickly develop a solution or response based on his experience. This can manifest itself as perception bias when SMEs use aspects of the Recognition-Primed Decision (RPD) pattern matching process [KLEI 01].²⁸ Such bias may not be wise to mitigate. However, until one can identify, measure, and mitigate perception bias, we have little understanding of practical bias. *Practical bias* is not a category or subcategory of bias. It is a measure of the magnitude and importance of the impact of participant inconsistency and inaccuracy. In other words, how much does bias skew results.

3. Performance Appraisal

Supervisors have used many methods to evaluate human performance over the years. Some of these means are purely qualitative in nature. Methods that describe the performance without ranking performance against others are known as absolute rating systems. There are four general methods involving absolute rating systems: behavioral checklists, essays, critical incidents and graphics rating systems. *Behavioral checklists* are similar to declarative knowledge in that they merely state facts regarding the existence or non-existence of a behavioral trait. These checklists are Go/No-Go in nature

²⁸ The RPD model is described in subsection II.G. Naturalistic Decision-Making.

and fail to indicate a level of performance. *Essays* allow raters to provide a more extensive description of the observed performance without limiting the assessment to a specific list of behaviors. However, essays do not provide standard rater responses and require a great deal of time to complete. *Critical incident reports* provide specific examples of performance, but require raters to witness the act [CASC 98]. Thus, essays and critical incident reports typically concentrate on procedural knowledge by allowing the rater to place the facts in context of the situation in which they were performed.

In an attempt to provide a quantitative means of assessing performance, supervisors can use *graphic rating scales*. These scales consist of a series of performance-based questions with standardized scales for evaluators to provide their assessment of subordinate behavior [CASC 98]. One example of a graphic rating scale is a Likert Scale. Likert Scales have an odd number of possible responses with one side of the midpoint representing substandard performance and the other side of the midpoint representing above average performance. The midpoint represents average performance. Scale values are general and subjective in nature but provide a means of quantifying subordinate performance. Examples of possible responses equated to a 5-Point Likert Scale are outstanding, above average, average, below average, and poor.

Graphic rating scales provide evaluators with four advantages over using open-ended questionnaires. First, graphic rating scales require less time to complete since they only require evaluators to choose one of the available options. Second, they allow evaluators a means of converting qualitative information into quantitative data. Next, since they are less time consuming, assessment forms can include more questions allowing for a broader assessment of an employee's performance. Finally, quantive employee performance data allows for comparison across evaluators and evaluates. Thus, graphic rating scales help evaluators capture aspects of procedural knowledge of individual behavior by acquiring more information about the employee while converting qualitative information into declarative knowledge.

Understanding bias is present in the assessment of human performance, Smith and Kendall suggest human resource personnel can assist supervisors in assessment of personnel by providing better assessment worksheets. These researchers developed a

rating scale consisting of a series of assessment questions with possible responses which include explicit examples of performance for each response listed [SMIT 63]. This scale is often referred to as the *Behavioral Anchored Rating System*.

Creation and validation of such evaluation forms is expensive and time consuming. However, they provide supervisors with a powerful yet relatively simplistic tool to assess the performance of their subordinates. More complex and time-consuming assessment methodologies have been devised to provide a better assessment of personnel performance. According to King et al., over time, the Behavioral Anchored Rating System has proven itself as viable and reliable an assessment process as systems that are more complex [KING 80].

The *behavior observation scale* is a hybrid version of a graphic rating scale and behavioral check lists. The scale allows the supervisor to track the frequency of specified occupational behaviors [TZIN 00]. Because of this, it provides more information about the kinds of behavior a subordinate is performing, but still fails to address the quality or context of this behavior.

The most often used method of assessment is the graphic rating scale [CASC 98]. Each performance appraisal technique is subject to the observation and judgments of the supervisor. As such, they are subject to misinterpretation and bias. Some performance appraisal techniques are better at mitigating misinterpretation and bias than others.

G. NATURALISTIC DECISION-MAKING

Klein characterizes *naturalistic decision-making* (NDM) as a paradigm designed to describe how people perform rather than being a method to improve performance [KLEI 97]. The focus is on how experts use their experience to make decisions when concerned with the execution of tasks in complex environments [ZSAM 97]. Cognitive psychologists have demonstrated that, for expert decision makers, methods and models associated with NDM more accurately describe the human decision-making process than previous paradigms. This is especially true when the situation involves a “high stakes, dynamically changing environment, time pressure, (with) ambiguous or incomplete goals” [TOLK 02]. These characteristics typify decisions made by military personnel during times of crises decision-making and execution of military operations.

In the late 1980s, Klein developed a theoretical model of decision-making referred to as the Recognition-Primed Decision (RPD) model. The *RPD* model asserts that expert decision makers use pattern matching to provide viable solutions to a situation. When an expert cannot match the situation to a known pattern, he uses a modified decision-making process to provide a solution until the situation changes. In these situations, the expert may modify his mental model of the world or generate a story to explain the difference in what he is observing and what his mental model tells him should be occurring. Research has validated the RPD theoretical model as a decision model offering merit for military operations. However, as of January 2004, no computational implementation of the RPD model at the operational-level for military decision-making exists [KLEI 01]. RPD was never meant to be a computational model with predictive capabilities. It was developed to help understand how expert decision makers draw conclusions and select a course of action.

As with any model, RPD has its limitations. Due to the Law of Small Numbers, using RPD, or any model, for describing the decision-making process has limited statistical strength if one has a limited number of SMEs.²⁹ This could lead to an incomplete assessment of the decision-making process. Also, using experts exposes the process to human error. Although less likely than non-experts, SMEs may introduce bias into the decision-making process by negating plausible courses of action due to their incomplete collection of situational patterns. This bias comes in the form of knowledge-based mistakes, decision errors, and judgment errors.³⁰ Thus, even though “the decision processes typically studied in NDM consist of a series of decisions or a sequence of intermediate outcomes,” validating agents must use it with care to limit possible negative effects from potential SME bias [LIPS 97a]. Nonetheless, the nature of the validation process for HBR models, where one must take into account the context in which the task is being performed, suggests a fit between the face validation process and the NDM paradigm.

²⁹ The Law of Small Numbers takes effect when a person over infers the likelihood of the frequency of an event based on a limited number of observations [TREV 71].

³⁰ “Decision errors pertain to situational assessment, mental models, and sequential option generation/evaluation rather than concurrent choice” [LIPS 97a].

The NDM paradigm is applicable beyond the collection of referents and the face validation of HBR models. Validating agents can also apply its context dependent nature to the training and retraining of SMEs for the validation process [COHE 97] [LIPS 97b]. Validating agents must train and focus SMEs to ensure SMEs only assess the model for the specific domain. If problems occur with performance of the SME that require retraining, remedial training methods must also be domain specific [LIPS 97b].

Since face validation concerns experts making decisions about performance, it is apparent that the NDM paradigm is applicable to the face validation process where an assessment of the model's performance is made for a specific yet still complex environment. Specifically, validating agents may use the RPD conceptual model to validate HBR models and to train SMEs to perform validation for combat tasks through pattern matching.

Methods used by NDM researchers, such as cognitive task analysis (CTA), have been used for the initial stage of simulation design to assist in identifying important aspects of the task to be modeled [MILL 97]. This technique has similar requirements to validation techniques which require SMEs to assess a model in a context dependent situation. However, CTA requires one to look deeper than just the overt behaviors of a decisions maker.

Klein defines a task analysis as the direct observation of a person performing an action resulting in a detailed description of the tasks one accomplishes in order to achieve a goal. A *cognitive task analysis* is a more extensive/detailed look at cognitive components of the task. It seeks to describe the cognitive processes underling the performance of tasks and the cognitive skills required to respond appropriately to complex situations [KLEI 00]. Thus, it examines actions and the decisions leading to those actions.

A CTA does not predict actions. Information collected by performing a CTA can be used to produce a descriptive model developed through interviews with SMEs and is

qualitative in nature. In the past, CTA studies have been conducted for the design of human-computer interfaces, instruction and training, organizational design, system development, product design and marketing.

Many variations of CTA have been developed. Klein describes CTA as consisting of five steps: identifying sources of expertise, assessing the knowledge, extracting the knowledge, codifying the knowledge, and applying the knowledge [KLEI 00]. Aronson's taxonomy includes four phases: knowledge elicitation, analysis, knowledge representation, and validation [ARON 02]. Finally, Harvey separates the process into four phases: preliminary phase, identifying knowledge representation, knowledge elicitation techniques, and representations [HARV 01].

Using Harvey's phases, the *preliminary phase* requires individual(s) performing CTA to become conversant in the area they wish to study. It may consist of reading relevant professional or training manuals, unstructured interviews with SMEs, and participant questionnaires to collect information about the tasks required to achieve a goal or accomplish a task [HARV 01].

After achieving a sufficient understanding of the basic issues and tasks relevant to the problem domain, the next step is to determine how best to represent knowledge. Two ways of representing the knowledge are procedural and declarative. The factual or conceptual nature of declarative knowledge allows one to use the information in ways not originally foreseen. Since procedural knowledge is a more precise means of describing how an individual accomplishes a task, it is an efficient but less germane means of depicting how to perform a task. When determining which data representation to use, the individual(s) conducting the CTA must consider the nature of the information and processes to be modeled [HARV 01] [WRAY 92].

With a basic knowledge of the problem space and a decision on how to represent the domain knowledge determined, collection of the detailed knowledge set is undertaken. Data collectors usually conduct this phase using structured interviews of SMEs to gather significant content that researchers will analyze and model developers will codify [HARV 01].

Information representations can take many forms (e.g., flow charts, structured English syntax, entity relationship diagrams, Unified Modeling Language (UML) diagrams, etc.) [HARV 01]. There is no prescribed format for representing the information gathered during a CTA. The specific purpose of the CTA and the complexity of the tasks one is modeling will steer the individual(s) conducting the CTA to choose one or more of these methods for representing data. The more complex the task, the more important it is to have a well-understood language or technique for representing the information collected.

H. ASSESSMENT OF PREVIOUS WORK

Pew et al.'s statement that "few individual combatant or unit-level models in the military context have been validated using statistical comparisons for predication" points to a major issue with emergent military simulations [PEW 98]. Until recently, a limited number of research efforts have attempted to address the issue of validating HBR models. Some of these most prominent have been project Agent-based Modeling and Behavior Representation (AMBR), Birta and Özmira's automated result validation model, Caughlin's metamodel methodology, Gonzalez and Murillo's validation through automated observations, and current work on alternative scales for face validation results [AIR 01] [BIRTA 96] [CAUG 95] [GONZ 98] [HARM 03]. Additional work such as Tactical Decision-making Under Stress (TADMUS), demonstrated insights to issues such as SME bias [BARN 93] [HUTC 96a] [HUTC 96b].

Project Agent-based Modeling and Behavior Representation (AMBR) is an Air Force Research Laboratory (AFRL) program designed to "advance the state-of-the-art in cognitive and behavioral modeling for military applications" [AIR 01]. Researchers compared and contrasted HBR architecture implementations as they performed a series of "standard problems" in a simulated environment. During the project's initial phase, program personnel conducted a comparison of the effectiveness of four cognitive architectures: ACT-R, D-COG, EPIC-Soar, and iGEN.

An impartial moderator, BBN Technologies (<http://www.bbn.com/>), handled the comparison of the models and completed the study in 2000. The focus of the initial phase was multi-tasking. The domain was a simplified version of an enroute air traffic control system. Model developers modified and integrated each cognitive architecture into the

virtual air traffic control system and exercised the architectures to determine their ability to simulate the behaviors and perform in a multi-tasking mode. All the models were able to replicate the referent within tolerances. Experimental control personnel noted the differences in how each architecture implemented the multi-tasking requirement.

BBN Technologies' review of the methodology used during the study identified many important issues. Two major criticisms were the limited number of tasks and sparse number of referents used during the comparison. These issues made it difficult to perform an exhaustive comparison of the capabilities of the cognitive models. The referent used in the study also lacked the ability to make a "head-to-head comparison" of the models. Due to limited time for coding modifications, the architecture implementations lacked the capability to represent expert cognitive processes [GRAY 00].

A summary of the results of the study by BBN Technologies indicates the focus of the project was too vague. Were they to compare the overt behaviors of the models or the cognitive process behind the actions? Were the architectures supposed to simulate behaviors at the performance level or at all levels of interaction [GRAY 00]? These questions reflect the difficulties of comparing the capabilities of cognitive models. They also identify problems with a lack of consistent validation standards for HBR models.

Although phase one of Project AMBR failed to provide a comprehensive comparison of the four initial cognitive models, it did help to identify some of the fundamental difficulties with such a process. Although its focus was narrow, a specific non-real world task with limited referent, it is a starting point for future work in the development of cognitive model comparisons.

In 1995, Caughlin introduced the idea of using reduced order metamodels to validate models and simulations. He claimed this new method would be a more timely and cost effective means of validation.

The creation of a metamodel requires *a priori* knowledge, data, metamodel structures, and rules to determine which original model will produce the referent [CAUG 95].³¹ Caughlin describes two methods researchers can use to construct metamodels for validation, direct and inverse (Figure 9). The direct method requires creation of a second

³¹ *A priori* knowledge is knowledge derived "independent of all particular experiences" [ENCY 02].

model, the metamodel, composed of subcomponent models that are lower fidelity replicas of the original components. The issue with the new, lower-fidelity metamodels is the difficulty of ensuring they properly represent the original model and all its functionality. Traceability of the direct method is less of an issue with the inverse method. The inverse method produces a reduced order model using input data and output results from the original model. Although a mathematical approximation of the initial model, the metamodel created using the inverse model, has to deal with issues relating to fidelity, sensitivity, and accuracy of results [CAUG 95].

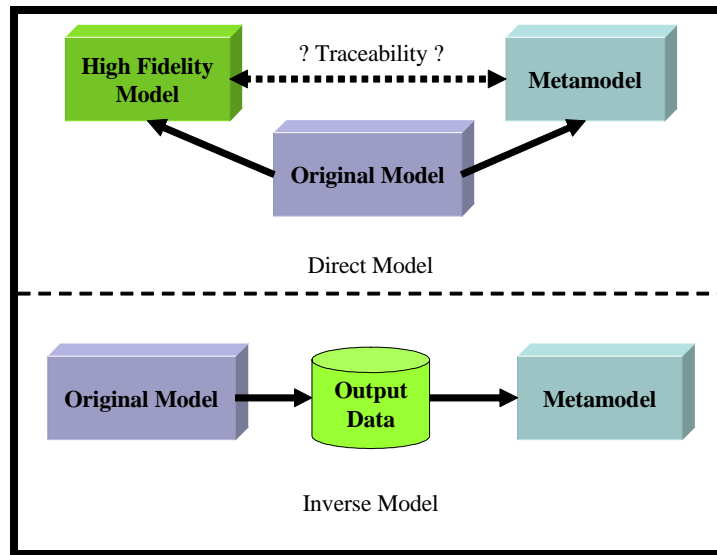


Figure 9. Metamodel Correspondence From [CAUG 95]

Caughlin’s metamodel approach to validation holds promise for analytical models that can be reduced to a more simplistic representation. However, this method of validation is not applicable to analytical models that are already in their most simplistic state. Nor has anyone shown the method to be applicable to models whose complexities make it impossible to create metamodels (e.g. cognitive models).

Birta and Özmırak proposed an automatic means to uniformly “validate” discrete, continuous, and combined simulation [BIRT 96]. Their technique focuses on an automated *face validation* of a model.³² They felt a single face validation of a model

³² Birta and Özmırak used the term “*behavioral validation*” in their paper. Although not specifically defined the technique is similar to face validation. To reduce confusion the term face validation is used in the section as a replacement for the term behavioral validation. It is NOT restricted to the validation of human behaviors.

could not perform an “absolute” validation. Instead, an experimental process is required. Figure 10 shows the four modules contained in their process: simulation model, validation knowledge base, experiment generator, and evaluator.

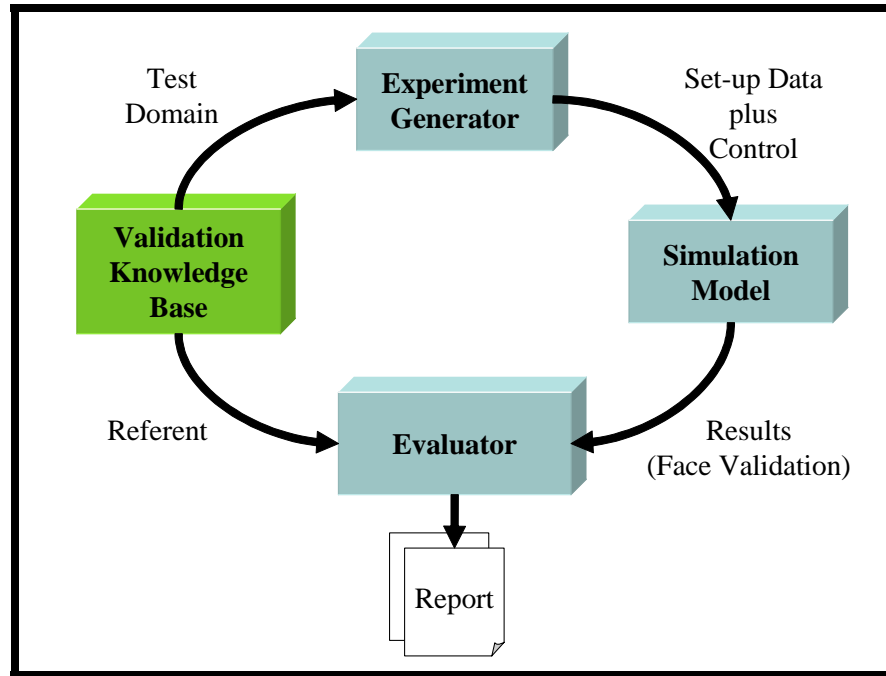


Figure 10. Global Architecture for Birta and Özmırak's Automated Result Validation Model From [BIRT 96]³³

The *simulation model* is the implemented program representing the system the user wishes to simulate. *Validation knowledge base* (VKB), the key component of the model, is the fundamental knowledge of input and associated outputs for the model. It represents the referent required by the model to meet its design specifications and intended use. Researchers use the VKB to develop the experiments used to validate the model's performance and the data to compare with the model's results. The *experiment generator* uses the input values provided by the VKB to design test cases for the simulation. Its goal is to produce the minimum number of test cases required to ensure a comprehensive validation of the model. Finally, *evaluator* takes the results from the

³³ Birta and Özmırak used the terms Reference Data and Behavior Data. These terms are changed to Referent and Results, respectively, to make them consistent with the terminology of this document.

simulation runs and compares them with the referent provided by the VKB, conducting a “critical evaluation of the simulation model output” [BIRT 96]. The results of the comparison are stored in the *report* files.

Birta and Özmırak use dynamic objects to identify the data required by the VKB. The dynamic objects are abstractions of dynamic behaviors represented in the simulation. A dynamic object, O , is described as an ordered pair of vectors X and Y where $O = (X, Y)$. X is the generalized input and Y is the output of the object. A causal relationship existing between the two vectors infers a change in X results in a change in Y . The fundamental property of all dynamic objects is their “ability to generate (exhibit) behavior over some prescribed time interval” [BIRT 96].

The VKB must possess all possible instances of the dynamic object. This means an exhaustive search of the problem space must occur to ensure every possible X , Y combination for the dynamic object is represented in the VKB. These pairings are a set of three disjointed types of specifications: formal, qualitative, and observable.

Formal specifications are X , Y relationships that always hold true (e.g., a 70-ton tank weighs more than a 60-ton tank). A qualitative specification displays the causal relationships between the input and output vectors (e.g. the main gun of a tank stops firing when it is out of ammunition). Finally, an observable specification is a means of ensuring the simulation replicates real-world behaviors when the experimental generator presents similar situations. This data is derived from the observation of previously validated simulations or real-world systems [BIRT 96].

Birta and Özmırak’s knowledge-base approach to model validation is a means of face validation. It attempts to accomplish validation through an automated system. This can reduce the bias injected into the face validation process by SMEs. The VKB appears to be a set of all available referents, powerful in its content but unlikely to be exhaustive for topics such as human behaviors. The approach also fails to address the non-deterministic nature of human behaviors.

In 1998, Gonzalez and Murillo proposed a method to validate human behavior models by means of automated observation. The technique allows a human behavior model to watch and learn from SMEs performing procedures in a standalone or

networked simulation. Computerized agents compare the behaviors of SMEs and simulations performing the same tasks to determine if the model's actions were similar. Later, additional SMEs can analyze the differences noted by the computerized agents to determine if the simulated behaviors were viable [GONZ 98].

Another aspect of this method is its ability to allow models to learn from SMEs as the two execute in parallel environments. As "serious" inconsistencies arise between the actions of SMEs and the simulation, a difference analysis engine (DAE) compares the two actions. If both actions were viable, the DAE would note the differences and allow the simulation to continue. If the computerized agents judge the model's behavior to be inappropriate, the automated system modifies the model's behavior to match the performance of the SME [GONZ 98]. This is similar to the training of a neural-network. It is also limited to the extent of modifications it can make based on the type and amount of input data available and the parameters of the algorithms.

Although the methodology may provide a means of training models, it must still address the issue of training behaviors valid for a simulation environment instead of replicating human behaviors in the real world. Developers face the same problem when using the method to validate simulation behaviors. Do these actions/behaviors transfer to the real world? Furthermore, the problem of creating a deterministic program to assess a non-deterministic model of behaviors demonstrating a non-linear nature is NP-complete and thus computationally intractable [MALL 88]. The method is another means of conducting a face validation of a simulation; however, as of January 2004, it has not been prototyped and tested.

The Defense Modeling and Simulation Office has determined that the current VV&A process for HBR models is inadequate. Work currently underway by Harmon and Metz seeks to determine if HBR model validation can be broken down into a series of validation levels based on the quantitative nature of the information available to assess them versus the current subjective methods [HARM 03]. Preliminary results from this research are due the summer of 2004.

Goerger presents an alternative methodology, which uses a continuous scale for validating HBR models instead of a binary valid/invalid scale [GOER 02]. The scale is

anchored on one end by a simple reactive agent HBR model and on the other end by the optimal HBR model, a human being. A model can be placed along the continuum of the validation scale indicating its degree of validity and allowing a relative comparison of similar models. The author's methodology addresses the diversity of HBR models and the varying degrees of information available to validating agents based on the model representation utilized to codify the theoretical model. Goerger argues that a validating agent can provide a more extensive assessment of a model's capabilities if the agent can query the model's cognitive process for information on its situational awareness and the plausible courses of action it is considering. With this information, the validating agent can assess if there are issues with the development of an adequate situational awareness, the cognitive process, or if the model lacks the diversity of options to address the situation. The methodology fails to address the

The *Tactical Decision Making Under Stress* (TADMUS) program developed a decision support system for enhancing the quality of the air warfare decision-making process. Aegis ship commanding officers and tactical action officers engaged in demanding littoral scenarios using a mock up of their current Aegis displays and performance was recorded. These scenarios were characterized as involving time-sensitive, ambiguous, dynamic situations. Significant improvements in air warfare decision-making performance (i.e., improved situational awareness, more of the correct tactical actions were taken, and decreased levels of communications) resulted when decision makers used the new decision support system [BARN 93] [PERR 93] [HUTC 96a] [HUTC 96b].

One separate, but related, issue investigated under the TADMUS program was cognitive bias in the decision-making process. Tactical action officers engaged in challenging scenarios and performance was recorded and analyzed. Biases in the air warfare decision-making process were identified; these biases included anchoring, contrast and confirmation [BARN 93] [PERR 93].

III. METHODOLOGY AND EXPERIMENTAL DESIGN

The methodology for modeling human-based representations draws upon information from three distinct yet related fields: modeling and simulations, human behavior representation, and behavioral and cognitive psychology. Illustrated below in Figure 11, each discipline has a unique perspective on how its addresses aspects of creating viable HBR models that, until recently, had little in common with the other two disciplines. When considered as a whole, there are key elements from each discipline common to these domains. The common area of interest for this dissertation is represented by the intersection of the overlapping ovals in Figure 11.

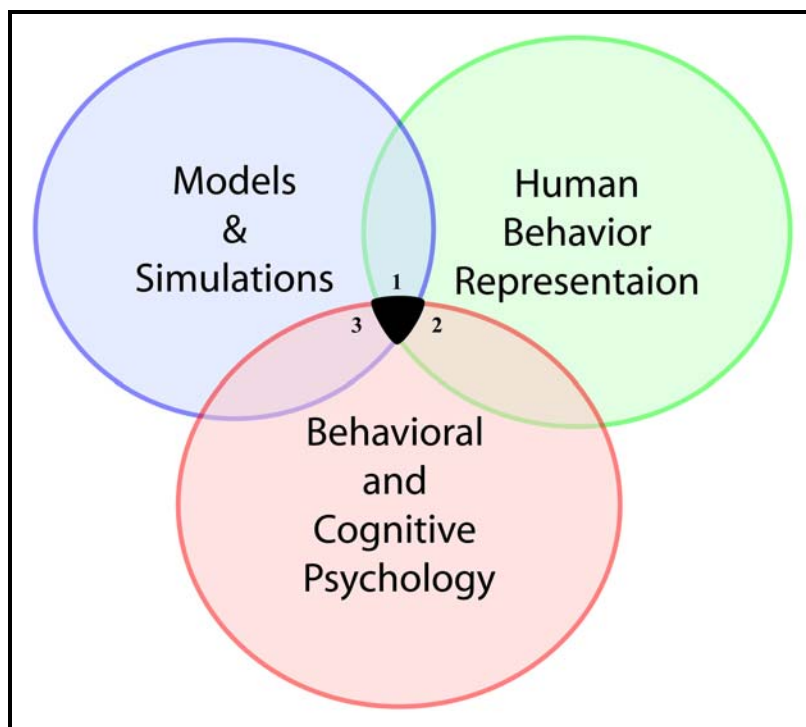


Figure 11. Area of Interest for Validating Human Behavior Representations Models

For decades the DoD M&S community has built, validated, and used physics-based models. As noted previously, the literature contains very few references to formal research to address the issues involved in creating, implementing, validating, and using HBR models. Initially using rule-based models to implement human behavior, DoD M&S began integrating more complex HBR models into simulations in order to study more advanced concepts and requirements, research, development, and acquisition of

weapons systems, and to enhance training for military personnel. The community has discovered that validation procedures for physics-based models are not adequate for HBR models [DEPA 01f].

For over a century, the cognitive and behavioral psychology communities have theorized about, and attempted to model, human mental processes and behaviors. Unlike most physics-based models, human behavior models are not mathematically-based, thus making it difficult, if not impossible, to codify these models.³⁴ However, human behavior research has collected vast amounts of data that is available to modify and validate human behavior models. Although this data may not be directly applicable for use as referent, it can help identify the type of data to be collected for use in developing, validating, and using a specific HBR model.

The body of HBR research documents numerous theoretical models, many of which were developed by psychologists to understand, explain, and predict human behavior. Most models are still in the conceptual development stage and not implemented in computer simulation. Some models lack sound psychology underpinnings; others lack a mathematical implementation or viable referent to allow them to be codified. Based on today's VV&A standards, the vast majority of these models are not acceptable for use in DoD simulations.

An important area of research involves bridging the gap between theoretical models of human behavior representation developed by the cognitive and behavioral psychology communities and the DoD M&S community's models and simulations. Drawing on the strengths of these communities may lead to the development of human behavior models with viable referents that can be implemented in DoD models and simulations, and properly validated.

³⁴ Mathematically-based models are those which have a direct relationship between the parameters, input-data, and results (e.g. if force equals mass times acceleration, then knowing force and acceleration, one can calculate the mass). Models which are not mathematically based are qualitative, have unknown parameters, or lack data which can influence the results. For example, a decision maker's memory capacity, training, experience, and stress level may influence his capability to identify the situation and select an appropriate action. Not knowing the relationship between these factors makes it impossible to predict the action the decision maker may select.

One issue that arises with attempts to mathematically model human behavior is the need to do so within the context of understanding and addressing the cognitive processes of the implementation. Today, this would require utilizing evaluation techniques, such as introspection, which have not proven to be viable for the assessment of real-world and simulated behaviors. Another issue is designing an appropriate experiment to collect the referent used for the development of the theoretical and implementation model. Normally this is accomplished with use of SMEs who, as human beings, may bias the information they collect. Similarly, bias may also be exhibited by SMEs who assess the face validity of a theoretical model or its instantiation.

The intersection of the overlapping ovals in Figure 11 is bounded by M&S HBR model requirements and implementations, referent and theoretical models provided by the psychology community, and mathematically based HBR models. Near the psychological community and nexus edge (Figure 11, Edge 1) lie theoretical HBR models described by paradigms such as NDM. Also along this edge are referent and theoretical models of emergent and predictive behavior for specific groups of individuals such as terrorists. Along the nexus and M&S boundary (Figure 11, Edge 2) the requirements for new HBR model implementations that address the requirements of future models and simulations intended to meet DoD training and analytical needs. Along the final edge of the overlapping area (Figure 11, Edge 3) the mathematical models of HBR which replicate the physical constraints of human performance: stress models, sleep deprivation models, etc.

Within the boundaries of the intersection of the three domains reside M&S HBR models such as ACT-R, Soar, COGNET, etc. and referents for validating and running HBR model implementations. Also within the nexus are numerous areas of interest requiring further research to clarify their influence or composition. These include the type of referent to build and validate HBR models, the development of theoretical models that can be codified, the strength and weakness of current HBR model architectures etc. One common issue in this generally uncultivated area of research is the role and influence of SMEs.

SMEs exert influence in identifying model requirements, the collection of referent, the validation of theoretical models, and the validation of model implementations. Their influence can be affected by many factors. Factors that exert a systematic influence altering SMEs' observations are called biases. Little is known about the influence of biases on SMEs' products for this field. If biases exist, what are they? How can they be identified? Can the magnitude of their impact be measured? Does their effect play a negative or positive role in the products produced? Can their influence be mitigated? There are numerous other research questions residing within the junction of the three communities. This research attempts to address the presence of SME biases and their effect on the consistency and accuracy of face validation results for an HBR model implementation and lays the groundwork for future research.

A. SCOPE

A review of the literature reveals numerous papers and books on the validation of physics-based models, deterministic HBR models, and analysis and interpretation of validation results. However, the literature review did not surface any studies in the area of SME bias when conducting face validation of HBR models. The lack of fundamental research in this area raises four issues. One, the feasibility of using human behavior evaluation techniques to conduct face validation of HBR models. Two, the effect of scale on the validation of HBR models. Three, identifying and assessing the effect of SME bias on observed HBR performance. Four, identifying the effect of SME personality on the presence of traits and characteristics as a source of bias. These issues were selected as a place to beginning addressing the proper use of SMEs to provide a controlled assessment process for the validation of a human behavioral model implementation.

This research involves a series of studies designed to examine five questions.

- Can we identify SME bias in the assessment of simulated behaviors?
- Do SMEs assessing human behavior demonstrate the same types and amounts of bias as SMEs assessing simulated behavior?
- Do SMEs provide consistent and accurate assessments of behaviors viewed through a simulation interface?
- When biases are identified in the validation of an HBR model, does the removal of biased SME responses mitigate the effects of inconsistency and inaccuracy?

- Does the type of scale used have an effect on SME consistency and accuracy?

A series of studies were designed to investigate these questions by collecting data on SMEs using multiple assessment scales as they assessed human behaviors or simulated human behaviors. SMEs were focused on assessing individual soldier and squad leader performance of ground combat operations in an urban environment. The SMEs used referent derived from *FM 7-8: Infantry Rifle Platoon and Squad*, 2001, and *ARTEP 7-8-MTP: Mission Training Plan for the Infantry Rifle Platoon and Squad*, 2001 to assess individual behaviors. An example assessment worksheet used as referent for these studies is in Appendix A.

B. EXPERIMENTAL DESIGN

The purpose of this experiment is to investigate the aptitude of SMEs for assessing the face validity of an HBR model. A series of studies were performed as part of the experiment. The design was based on validation of MANA, an agent-based model, to collect data from SMEs. MANA provides the visual display of human behaviors for individual dismounted soldiers which are assessed by the SMEs for validity. To ensure the validation process utilized in the experiment was viable and consistent with current DoD validation procedures, the strategy was reviewed and approved as an acceptable and viable validation plan by members of the DoD verification, validation, and accreditation community (Appendix Q. Validation Plan).³⁵

MANA was chosen as the simulation for the validation exercise for several reasons. First, its ability to execute behaviors at the required entity level fidelity facilitated modeling behaviors for each soldier and civilian. Secondly, the area of operations could be limited to maintain SME focus on individual soldier and squad leader tasks. Additionally, SMEs had no prior experience with the simulation and therefore had

³⁵ One of those who reviewed the “Human Behavior Representation Validation Plan” for this dissertation was Dr. Dale Pace from the Johns Hopkins University Applied Physics Laboratory; a recognized leader in the commercial and DoD VV&A community. Dr. Pace had the following comment, “I commend you for the rigor you bring to both SME use and the HBR assessment in this plan. You employ appropriate methods in SME use” [PACE 04]. Other reviewers from the VV&A Community were Simone Youngblood, the DMSO VV&A Technical Director, Scott Harmon of ZETETIX and a member of the DMSO VV&A Technical Working Group, Susan Solick, Army M&S VV&A Standards Category Coordinator (SCC), and Marcy Stutzman from Northrop Grumman of the Navy VV&A Team.

limited preconceived opinions regarding its capabilities and limitations. Finally, MANA's simplicity in programming made development and modification of scenarios comparatively simple.

One set of measures of performance (MOPs) assesses the presence and effect of four SME biases: performance, anchoring, confirmation, and contrast.^{36 37} Additional MOPs identify and quantify consistency and accuracy. A third set of MOPs addresses the overall assessment level of the behaviors observed by SMEs. This research defines the MOPs as:

- *Performance bias* exists when a SME chooses not to provide definitive responses to 20% or more of the assessment questions.³⁸
- *Anchoring bias* occurs when a SME judges the first task and associated subtasks as a “Go” and, after viewing the second task and associated subtasks, judges the remainder of the model performance as “Go” for 90% or more of the assessment questions for which he provides a definitive response.³⁹ Anchoring bias is also evident when the SME judges the first scenario, associated tasks and subtasks as “No-Go” and then, after viewing the second scenario and associated subtasks, judges 90% or more of the assessment questions for which he provides a definitive response for the remainder of the model performance as “No-Go”.⁴⁰
- *Confirmation bias* exists when the differences in a SME's sublevel mean scores and level responses tend towards no difference in response or show a consistent difference in response and the overall response differs from this trend.⁴¹
- *Contrast bias* exists when a SME starts with a negative or positive assessment of the first task and, after viewing data differing from this initial opinion, the SME negates any further evidence in support of the original hypothesis and

³⁶ MOPs are quantitative measures or ranges of values [PIAN 01].

³⁷ Performance, anchoring, confirmation, and contrast bias are defined in II.F.2. Bias.

³⁸ A definitive response is a “Go” or “No-Go” assessment of the subtask, task, scenario, or overall assessment question. “Not Applicable” or “No Opinion” responses are not categorized as definitive responses.

³⁹ In accordance with doctrine, the squad failed to perform the second task and associated subtasks for “React to the Sniper Attack” to standard. The squad lost two personnel with the remainder of the squad not reacting to the sniper's attack or to the loss of personnel.

⁴⁰ In accordance with doctrine, the squad performs the second scenario and associated task and subtasks to standard. The squad successfully defended the building by destroying enemy forces attempting to seize the structure without the loss of any friendly forces.

⁴¹ Note: differences between sublevel mean scores and level responses may mitigate each other with the addition of more assessment responses; this does not indicate confirmation bias.

assesses the model based on the swing of opinion. In addition, the SME's accuracy data must indicate a shift in the accuracy trend from harsher to more lenient or more lenient to harsher as the assessment process proceeds. This shift occurs after the swing in raw score responses.

- *Inter-SME consistency* is achieved when 66.7% of SMEs' responses, when observing and assessing the same behavior, are the same: Go, No-Go, or Not Applicable.
- *Intra-SME consistency* is achieved when the raw score differences between a SME's mean sublevel assessment value and the SME's level response are less than +/- 0.5.
- *Intra-SME consistency impact* is zero or insignificant when the differences between a SME's mean sublevel assessment value and the SME's level assessment response do not change the relative value of assessment for the level (i.e. remains Go, No-Go, or Unknown).⁴²
- *Intra-SME accuracy* is achieved when the number of differences between a SME's assessment responses and the associated scale's key assessment values is less than 10% of the total number of assessed tasks when observing and assessing the same behavior.
- *Intra-SME accuracy impact* is zero or insignificant when the differences between a SME's assessment responses and the associated scale's key assessment values do not change the relative value of assessments (i.e. baseline scores are the same as the response scores: Go, No-Go, or Unknown).
- *Overall model validity score* is indicated by a cut off score 0.667. The MANA model is "valid" for use in modeling the proscribed tasks when the normalized mean scores from all SME responses to questions entitled Overall 1 and Overall 2 are equal to or greater than 0.667. Normalized values above this score fall into the range of responses, which SMEs were told, are "Go" or passing scores.

For purposes of this dissertation, SME consistence and SME accuracy are fundamental to understanding the impact of issues related to the use of SMEs. Text Box 1 provides the definition for SME consistency as used in this experiment.

SME Consistency – the ability to maintain logical correspondence between the average sublevel response score and the level score. In other words, deriving level responses logically/directly from sublevel responses.

Text Box 1. Subject Matter Expert Consistency Definition

⁴² Accuracy is calculated using an assessment key developed by a SME with access to appropriate training manuals and unlimited time to assess the model's performance. It is consistent at all levels and provides a relative reference point from which to compare SME responses.

SME consistency differs from consistency between SMEs, inter-SME consistency. Text Box 2 provides the definition for inter-SME consistency as used in this experiment.

Inter-SME Consistency – the agreement between SMEs on the Go/No-Go of a single behavior or the overall validity of the observed behaviors. The higher percentage of SMEs with the same Go/No-Go assessment, the greater the inter-SME consistency.

Text Box 2. Subject Matter Expert Consistency Definition

Text Box 3 provides the definition for SME accuracy as used in this experiment. The assessment key is a reasonable and consistent assessment of observed behaviors; however, it is not the only viable assessment.⁴³ It is a point of reference from which to compare SME responses. Any consistent, reasonable assessment of the observed behaviors would also be as suitable for use as the assessment key. SME accuracy scores would differ based on the assessment key used, but the overall effect would be similar. If an assessment key was consistent and unreasonable, statistically significant differences could exist between accuracy scores using the unreasonable assessment key and accuracy scores calculated in this research.

SME Accuracy – difference between the assessment key and the SME's assessment of each observation, where a difference is the assessment value from the key minus the assessment value of the SME for a given subtask, task, scenario, or overall question.

Text Box 3. Subject Matter Expert Accuracy Definition

1. Study Simulation

SMEs were asked to perform face validation of an HBR model embedded in a non-real-time, entity-level, constructive ground combat simulation (Figure 12). The reasons for using this type of system are: (1) potential bias introduced by human-in-the-loop simulations; (2) computational constraints of real-time systems; (3) physics

⁴³ In this context, reasonable means an assessment which does not assess obviously poor performance as acceptable or obviously proper performance as unacceptable.

limitations of games; (4) limited fidelity of aggregate level models; (5) the ability to precisely replicate scenario runs; and (6) ground combat domain complexities.⁴⁴

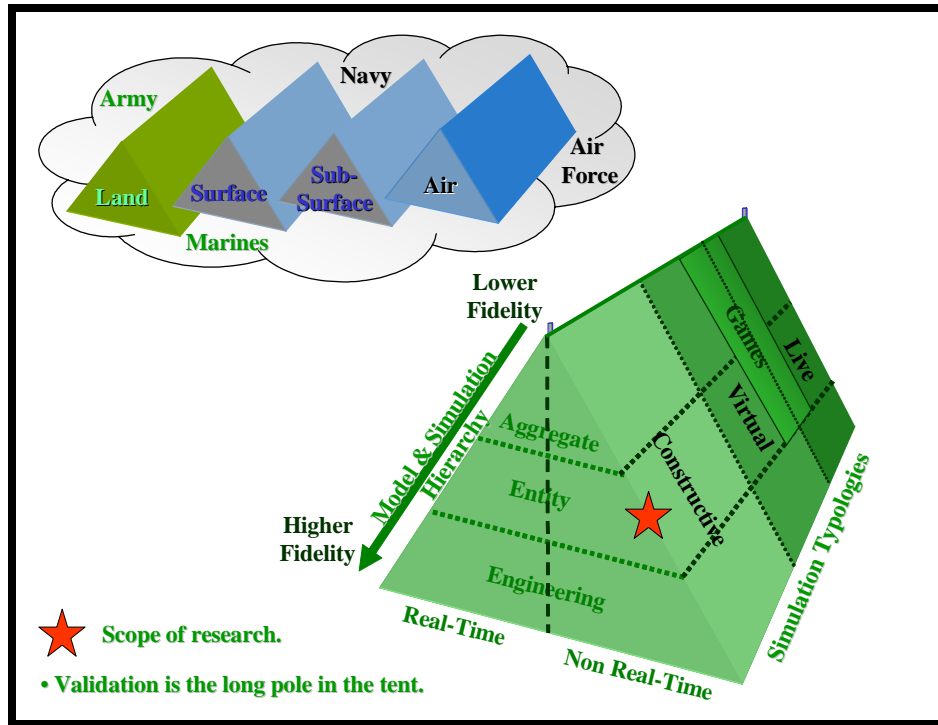


Figure 12. Model and Simulation Classifications From [⁴⁵]

2. Simulation Environment

The simulated training and practice environment was a road intersection with six buildings (Figure 13). The environment provided study SMEs with a visual display of MANA for practicing assessment procedures and to familiarize them with the assessment worksheet. This familiarization reduced the learning curve during the data collection phase of the experiment.⁴⁶

⁴⁴ Appendix Q discusses the simulation typologies, real-time vs. non real-time models and simulations, model and simulation hierarchy, military simulations vs. games, and model domains.

⁴⁵ See [HUGH '97][AMSO '00][LAIR '00][GOER '03].

⁴⁶ Appendix F. Experiment Environments and Scenarios describes the simulated practice environment, simulated McKenna environment, and urban scenarios in detail.

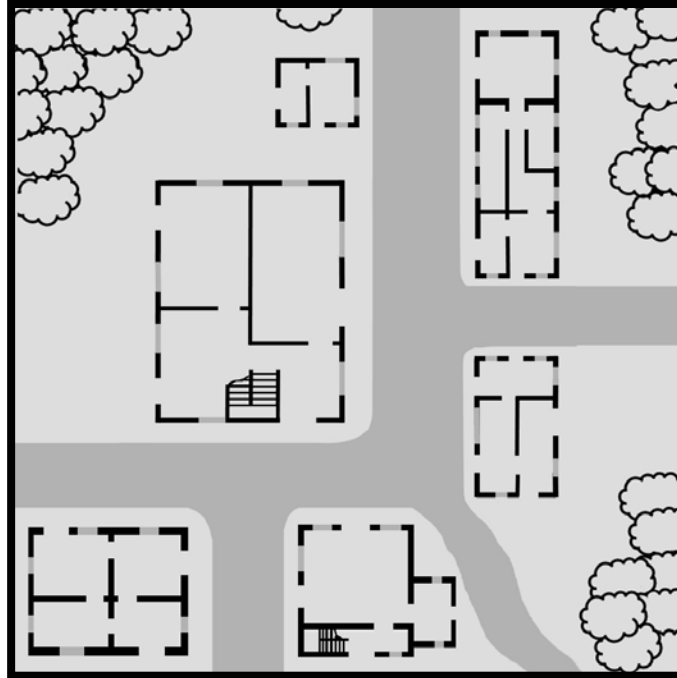


Figure 13. Training and Practice Environment Sketch

For the data collection phase of the experiment, research personnel used a second environment. A replication of the McKenna military operations in urban terrain (MOUT) Site, Fort Benning, GA (Figure 14) was modeled in MANA. This environment consisted of 28 buildings and supporting road network. The environment was selected for two reasons. First was the accessibility to data from experiments performed at McKenna such as the Natick study performed by Statkus, Sampson, and Woods in which they observed squad size units performing offensive and defensive tasks in an urban environment [STAT 03]. Second was a majority of the study SMEs were familiar with the McKenna environment.

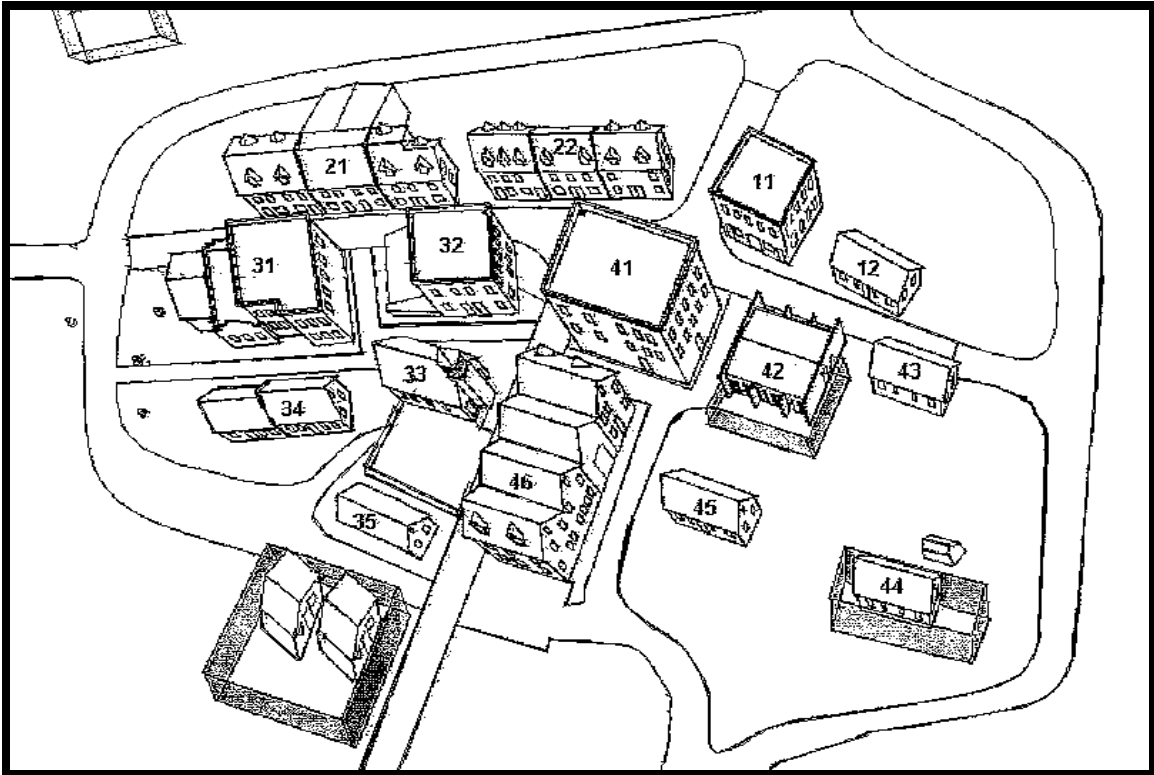


Figure 14. McKenna Test Environment Sketch From [STAT 03]

The data collection phase used the McKenna environment for all offensive and defensive test scenarios. While the offensive scenarios used the entire McKenna village as seen in Figure 14, the defensive scenario used only a portion of the south central section of the site. The defensive area encompassed building complex 46 (Figure 14) and segments of the adjacent buildings. Figure 15 shows the area of McKenna village used for the defensive environment.

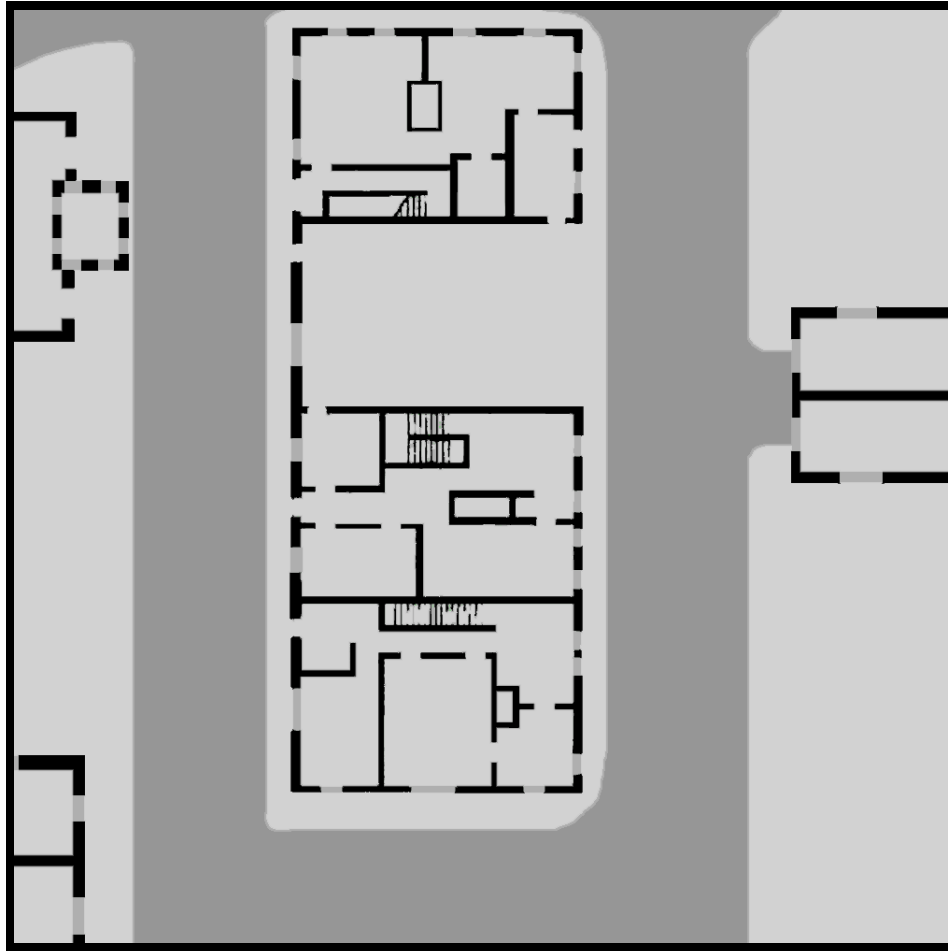


Figure 15. McKenna Test Environment Defensive Sketch

3. Scenarios

The SMEs reviewed four separate scenarios, one practice and three test scenarios. The practice scenario focused SMEs on defensive tasks and was executed in the practice environment. The scenario involved a squad-sized element of ten personnel defending a building against an attack by an enemy squad of nine personnel (Figure 16).

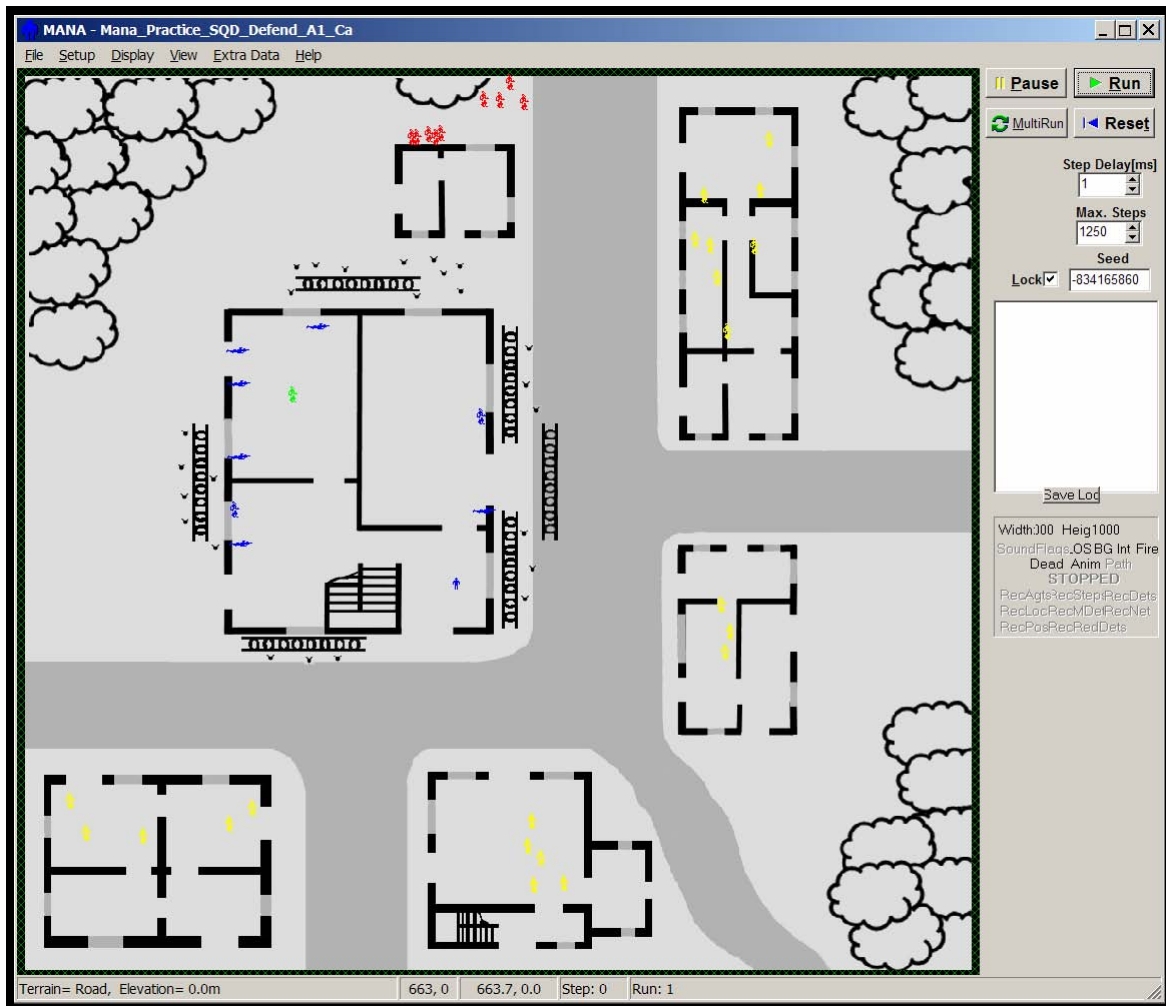


Figure 16. Training and Practice Environment Model Display and Scenario

MANA, the study model, ran the test scenarios over McKenna MOUT Site terrain. The test scenarios consist of two offensive scenarios and one defensive scenario. Both offensive scenarios placed enemy forces in accordance with the 2003 Statkus, et al. study and were modeled using all buildings in the McKenna MOUT Site. The first offensive scenario was scripted using data collected from Movement Scenario 1 of the Statkus study [STAT 03]. The second offensive scenario (Figure 17) was generated using the MANA simulation and designated waypoints. Both scenarios focused on a squad-sized element of nine personnel infiltrating the village of McKenna in order to rescue a prisoner of war (POW) from ten enemy personnel. The POW was located in a building to the south central portion of McKenna. The infiltrating squad could approach the village from any direction.

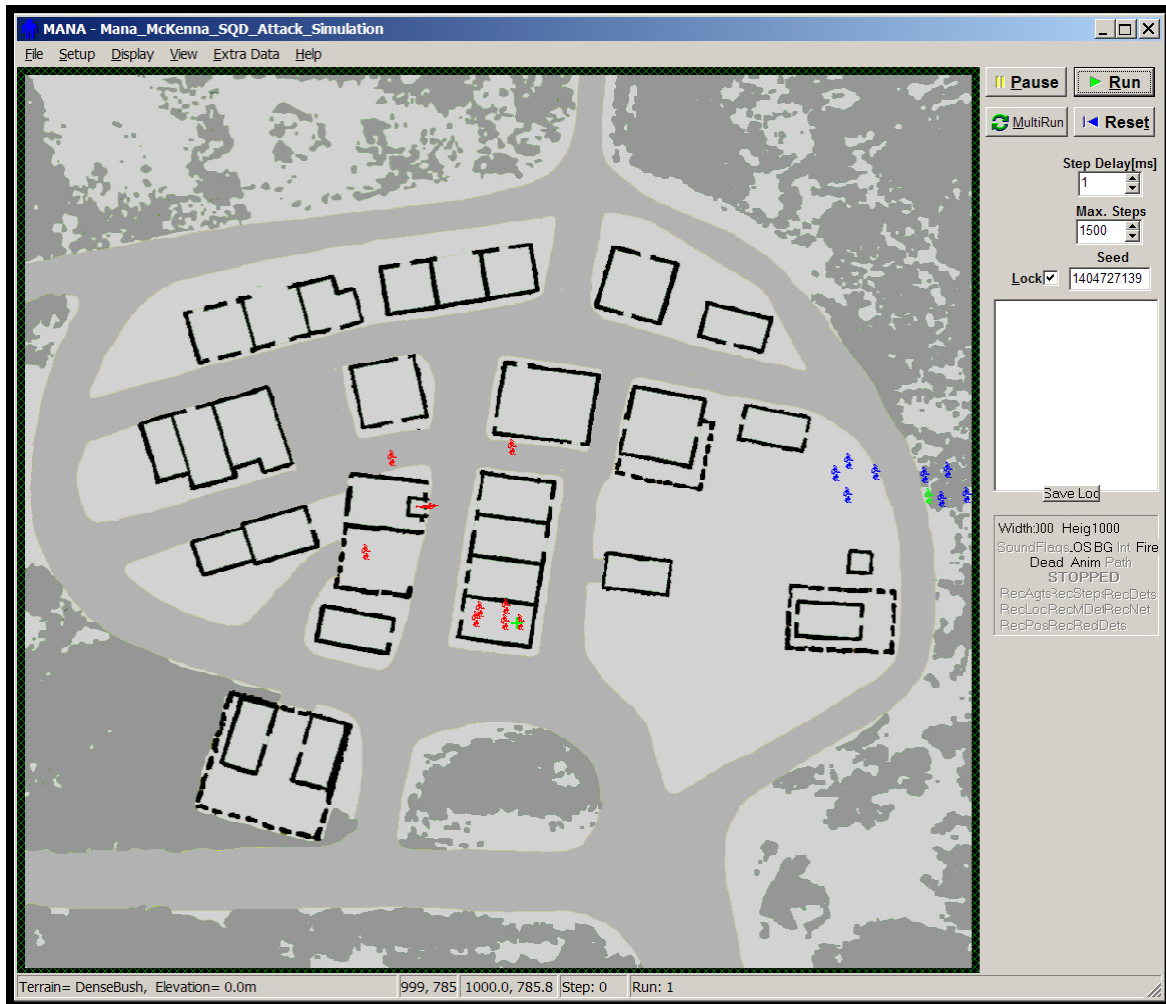


Figure 17. MANA Display of McKenna Test Environment, Offensive Scenario #2

The defensive scenario used the same general situation as the two offensive scenarios, but instead of assessing the action of the offensive operation, SMEs assessed the performance of the defenders. The defensive scenario focused on the actions directly surrounding the four building complex identified as **46** in Figure 14.

4. Data Collection Techniques

Demographic data was collected and the Neuroticism, Extraversion, and Openness Five-Factor Inventory (NEO-FFI) was administered. Demographic data included military experience, combat experience, video game and simulation experience,

and urban operations training.⁴⁷ To account for SME assessment response differences based on individual psychological makeup, the NEO-FFI was administered to all SMEs.

The NEO-FFI, describes personality based on an individual's reported pattern of enduring behaviors, emotions, and thoughts [COST 92]. These patterns define who we are and how we interact with others. The test assesses characteristics of an individual's personality based on measures of five broad personality factors: neuroticism, extraversion, openness, agreeableness, and conscientiousness.

In general, *neuroticism* is the measure of an individual's inclination toward negative emotions and instability, often indicating a limited ability to cope with adversity. *Extraversion* measures an individual's tendency for social interaction. *Openness* is a gauge of an individual's tolerance for new ideas and innovative ways of doing things. *Agreeableness* measures the individual's propensity away from animosity and towards empathy and sympathy for others. Finally, *conscientiousness* measures the degree of organization and inclination towards goal-oriented activities.

To make direct comparison of factors easier, raw scores for these five factors have T-Score transformations which distribute scores between 0 and 100 with 50 being the median score. Whether raw scores or T-Scores, the NEO-FFI values are normally distributed. Scores closer to the tails of the distribution imply a greater likelihood an individual will exhibit the indicated characteristics.

General characterizations can be made for the five traits. For neuroticism, individuals with higher scorers tend to be more anxious, apprehensive, and distressed while those with lower scores are likely to be more peaceful, self-confident, and resilient. Individuals with higher extraversion scorers are apt to be more outgoing, high-spirited, and talkative than individuals with lower scores who are prone to be more reclusive, repressed, and aloof. For the openness trait, those with higher scorers tend to be more imaginative, independent, inquisitive, and multifaceted compared to those with lower scores who are apt to be more traditional, less creative, and down to earth. Individuals with higher scores for agreeableness are likely to be more agreeable, understanding, tolerant, and polite whereas individuals with lower scores are prone to be more critical,

⁴⁷ Appendix G. Participant Demographics, Experience, and Training Questionnaire is an example of the demographics work sheet used in the experiments.

impolite, unforgiving, and unsympathetic. Finally, those individuals with higher conscientiousness scorers are apt to be more dependable, efficient, self-disciplined, and cautious where those with lower scores tend to be more undirected, disorganized, unreliable, and inattentive [COST 92] [COST 00] [COST 03] [MILL 04].

The NEO-FFI identifies and describes ten personality styles based on the pairwise interaction of the five NEO-FFI score categories. These ten areas are *Well-Being* (neuroticism and extraversion), *Defense* (neuroticism and openness), *Anger Control* (neuroticism and agreeableness), *Impulse Control* (neuroticism and conscientiousness), *Interests* (extraversion and openness), *Interactions* (extraversion and agreeableness), *Activity* (extraversion and conscientiousness), *Attitudes* (openness and agreeableness), *Learning* (openness and conscientiousness), and *Character* (agreeableness and conscientiousness) [COST 92]. Each personality style is broken down into four quadrants based on T-Scores greater and less than 50 for the first scoring category crossed with the T-Scores greater and less than 50 for the second scoring category.

SME assessment data was collected using worksheets modified from the *ARTEP 7-8-MTP* evaluation forms used for referent. As SMEs observed behaviors through the MANA interface, they recorded their opinions on the evaluation worksheets using a quantitative scale and by providing qualitative comments. Research personnel transferred the quantitative data from the assessment forms to Excel® spreadsheets. The Excel® spreadsheets were imported into JMP® for analysis.

Qualitative data extracted from the assessment worksheets was used for clarification or amplification of SME quantitative assessments. Information from the debriefing questionnaire was used to assist in modifying the experimental design for future experiments and to provide insight to possible issues with SME responses.

C. STUDY #1 EXPERIMENTAL DESIGN

This study was designed to investigate biases demonstrated by SMEs when responding to experimental scenarios based on their belief they were observing live or simulated performance. That is, the study investigated whether SMEs are susceptible to cognitive biases in the assessment of what they perceive as real-world data presented on a computer screen using a 2D map or textural display versus when they perceive they are assessing computer generated force (CGF) behaviors. Showing a similarity in anchoring,

contrast, and/or confirmation biases when assessing perceived CGF performance or perceived human performance will help determine whether SMEs apply the same criteria when assessing real-world performance and simulated performance under the same conditions. Recall the methodology used to rate performance in the experiment is based on the methodology used to evaluate live performance of soldiers.

1. Hypotheses

To determine whether SMEs demonstrated more, less, or equal levels of performance, anchoring, contrast, and confirmation bias when assessing perceived human performance as they do when assessing perceived simulated human behavior, this study was performed. The study was designed to identify and quantify the relative differences in biases between the two groups of SMEs, those who believe they are assessing simulated behaviors and those who believe they are assessing real-world behaviors, using the conventional human behavioral model validation methodology, face validation, for assessment. For the study, performance bias was measured by determining if a SME failed to respond to 20% or more of the assessment questions. Anchoring bias was measured by determining how far a SME varied from his initial hypothesis of the validity or non-validity of the model, regardless of the information presented when a mixture of proper and improper performance is present. Contrast bias was measured by determining if a SME rejected the hypothesis regardless of the evidence presented. Confirmation bias was measured by the amount a SME diverged from the hypothesis regardless of the evidence presented.

If there is no statistically significant difference in the levels of bias between the assessment of perceived human behaviors and computer generated behaviors for these measures of performance, then we conclude the methodology for assessing human performance is viable for assessing computer generated performance, or HBR models. The following is the null hypothesis for the first study.

H_0 : The assessment of human performance shows no difference with regards to bias for the two groups of SMEs using conventional validation methods as outlined in the Defense Modeling and Simulations Office (DMSO) Verification, Validation and Accreditation (VV&A) Recommended Practice Guide (RPG) for HBR.

Conversely, the alternative hypothesis is given as:

H_A : The assessment of human performance by SMEs shows differences with regard to bias for the two groups of SMEs.

2. Design

There were four dependant variables, also referred to as measures of performance, for this study: performance, anchoring, contrast, and confirmation biases. This research calculated performance, anchoring, contrast, and confirmation biases based on the formulas described in Chapter IV.C. Assessment. This determination was based on four distinct values, one per bias type.

The following are the independent variables for the first study. Primary factors are those taken into consideration. Secondary factors include independent variables that may influence the response variables but whose effect was not investigated in the analysis of the results.

- **Primary Factors**
 - a) Behavior Generator: human performance, model performance
 - b) Validation Methodology: conventional VV&A methodology
 - c) Subject Matter Expert(s): senior company grade Infantry officers
 - d) Cognitive Model: cognitive model architecture (agent, etc), level of implementation (squad level)
 - e) Domain: ground combat
 - f) Scenario: two attack and one defensive urban operation (Military Operations Urban Terrain; MOUT)
 - g) Simulation/Game: MANA
- **Secondary Factors**
 - a) Behavior Generator: None
 - b) Validation Methodology: None

- c) Subject Matter Expert(s): prior service and urban operations combat experience, MOS/Branch, years of service, experience, urban operations training, assignments, sex, national origin ^{48 49}
- d) Cognitive Model: cognitive model architecture (Neural-Network, Bayesian, MAS, etc), cognitive model representation (Soar, ACT-R, etc.), level of implementation (individual, team leader, unit commander, battalion commander, staff, etc)
- e) Domain: peacekeeping operations, air-to-air combat, air-to-ground combat, surface combat, subsurface combat, space, etc.
- f) Scenario: mounted infantry attack, mobile defense, static defense dismounted infantry attack, react to direct fire, react to indirect fire, defense in depth, movement to contact, etc
- g) Simulation/Game: JCATS, COMBAT^{XXI}, America's Army, CASTFOREM, OneSAF, JSAF, JWARS, etc.

The treatments are primary factors whose effects on dependant variables were measured. There were two levels based on SMEs belief of what generated the behaviors. Table 6 indicates the type and number of primary factors and levels. Of seven factors, only two have more than one level: behavior generator, and scenario.

Table 6. Study #1, Primary Factor Levels

Reference ID	Factor	Number of Levels
A	Behavior Generator	2
B	Validation Methodology	1
C	Subject Matter Expert	1
D	Model	1
E	Domain	1
F	Scenario	3
G	Simulation/ Game	1

⁴⁸ Combat experience includes peacekeeping, peace enforcement, or war.

⁴⁹ Assignments would be categorized by unit type e.g. Airborne, Air Assault, Light, Mechanized, Armor, Cavalry, Ranger, Special Forces, SBCT, etc.

Table 7 breaks out the design by SME group based on the factors simulation belief and scenario. The levels for simulation belief were human and simulated performance. The SME groups represent the study blocks.⁵⁰

Table 7. Study #1, Experimental Layout by Subject Matter Expert Group

SME Group	Offensive Scenario #1	Defensive Scenario	Offensive Scenario #2
1	<i>Human Performance</i>	<i>Human Performance</i>	<i>Human Performance</i>
2	<i>Simulated Performance</i>	<i>Simulated Performance</i>	<i>Simulated Performance</i>
3	<i>Human Performance</i>	<i>Human Performance</i>	<i>Simulated Performance</i>
4	<i>Simulated Performance</i>	<i>Simulated Performance</i>	<i>Human Performance</i>

3. Procedures

The study involved participants and the study team personnel. The study team personnel consisted of those individuals who executed the setup and data collection. The study participants, or SMEs, were mid-grade company level Army or Marine Corps officers who were enrolled in or who had completed their respective officer advance course and who had experience leading soldiers during military ground operations in urban and natural terrain.

Two versions of the study were conducted, one to refine procedures (*pilot study*) and a second for data collection (*base study*). The pilot study was performed prior to the base study and helped identify problems with experimental procedures, provided insight to additional data to be collected, improved data collection procedures, and refined the SME pool. The pilot study was conducted at the Naval Postgraduate School (NPS), Monterey, CA with a SME pool recruited from the Training and Doctrine Command Analysis Center (TRAC) - Monterey and NPS. SMEs were Captains (O3) or Majors (O4) with urban warfare training or who had completed the ICCC at Fort Benning or individual branch specific advanced course.

⁵⁰ A *block* is a group of equivalent research entities on which one trial of the treatment is measured [PETE 85].

The base study was designed to collect data regarding biases demonstrated by SMEs when assessing perceived human performance or simulated human performance displayed via computer simulation. The data collection portion of the study was conducted from 18 September 2003 through 24 September 2003 at Fort Benning, GA. SMEs were recruited from the ICCC student body consisting of senior First Lieutenants (1LT/O2) or junior Captains (CPT/O3) who have had urban warfare training and were in the ICCC at Fort Benning.

4. Set-Up

The study location was Room O-256, Infantry Captains Career Course (ICCC), Building #4, Fort Benning, GA. The facility was a 20 foot by 30 foot room, accommodating 13 tables large enough for SMEs with workspace for 10x14 workbooks, writing materials, and maps (Figure 18).⁵¹ The room had a 5 foot by 5 foot projection surface to the front of the SMEs. A video camera for recording research procedures resided at the back of the room. The facility had sufficient lighting for SMEs to record their assessments and electrical outlets for the laptop, projector, and camera system.

⁵¹ Appendix E. Experimental Material lists the materials required for this study.

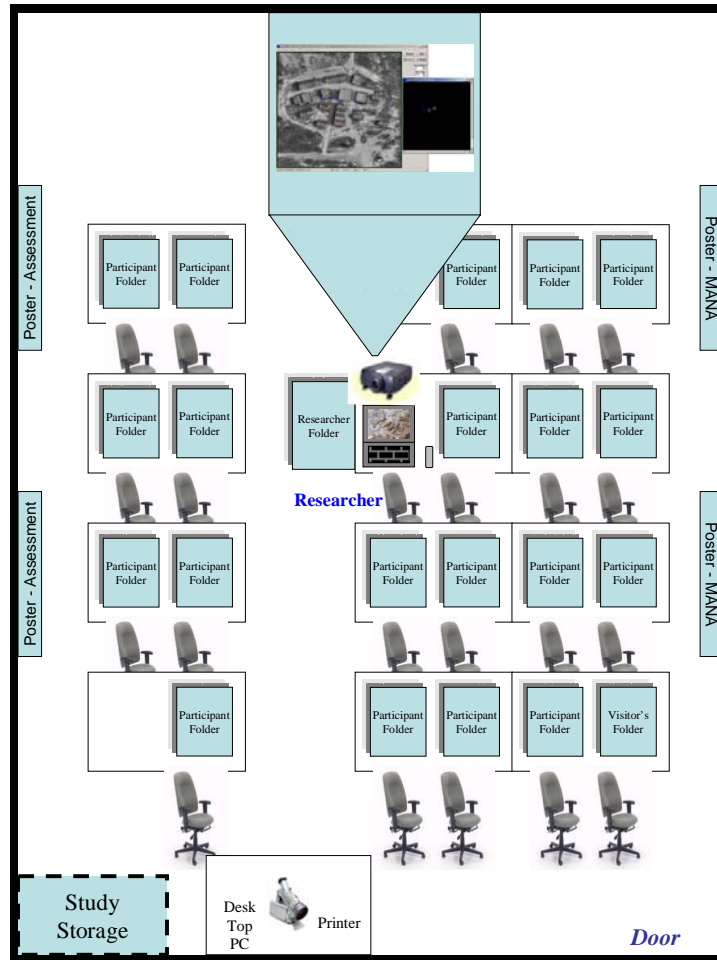


Figure 18. Room Layout for September Data Collection

Each workspace was set with a SME binder and blue pen. Spare materials (pens, worksheets, consent forms, etc.) were located at the back of the room on a research supply desk. Figure 19 is a picture of the room and its setup prior to the arrival of SMEs. The image was taken from the front row of tables.



Figure 19. Room Setup for September Data Collection

5. Study Phases

The study was conducted in five phases: in-processing, familiarization, training, data collection, and debrief.⁵² Upon arriving in the study room, SMEs underwent in-processing. This involved SMEs submitting a demographic survey (completed prior to arrival), reading and signing consent forms, and completing the NEO-FFI. The demographic survey was distributed to SMEs prior to the first day of the study to allow them to complete it at their leisure. The consent forms ensured SMEs understood the basics of the study and agree to the parameters of the study prior to participating.⁵³

During the familiarization phase, SMEs received a series of briefings designed to a) explain the objectives of the study, b) review appropriate doctrine and tactics,

⁵² Appendix B. Experimental Procedures outlines the timing and required materials for these phases.

⁵³ Appendix H. Consent Forms provides examples of SME consent forms.

techniques, & procedures (TTPs) for the task(s) to be assessed, c) acquaint SMEs with assessment techniques, and d) review assessment worksheets.⁵⁴

SMEs observed a training scenario in order to practice assessment procedures. The scenario was 22 seconds in length and displayed prototype examples of “poor” and “good” performance of defensive and offensive operations in an urban environment.

During the data collection phase, SMEs assessed performance for a series of tactical military tasks displayed through the MANA interface. The scenarios were 19 to 22 seconds in length and displayed “poor” and “good” performance of defensive and offensive tasks in an urban environment. Prior to the scenario run, SMEs sketched their interpretation of how the soon-to-be-assessed squad should approach the mission using a map and standard operational symbology. Prior to assessing performance, SMEs watched a fast forward run of the scenario to gain an understanding of how the scenario would progress. Next, a second run of the scenario was executed at a reduced speed with programmed pauses to allow SMEs time to record their observations. Each SME was asked to assess behaviors in two offensive and one defensive scenario. After all three scenarios were completed, SMEs conducted an overall assessment of performance and made an estimate regarding how the squad/model might perform if the scenario were executed in a different environment (e.g., jungle, arctic, desert, etc.).

At the conclusion of the study, each SME received a debrief that provided them with the results of their NEO-FFI, an exit questionnaire, and a one-page handout describing the study, its importance, and points of contact.

D. STUDY #2 EXPERIMENTAL DESIGN

This study was designed to investigate whether modified assessment techniques applied to cognitive model validation procedures mitigate the effect of bias on consistency and accuracy.

1. Hypotheses

To assess whether SMEs demonstrated more, less, or equal levels of consistency, consistency impact, accuracy, and accuracy impact when evaluating perceived human performance as they do when validating perceived simulated human behavior, a set of

⁵⁴Appendix K. Assessment Worksheets provides an example assessment worksheet.

studies were performed. They were designed to identify and quantify the relative difference in inter-SME consistency, intra-SME consistency, intra-SME consistency impact, intra-SME accuracy, and intra-SME accuracy impact for SMEs assessing human performance and simulated human behavior using modified assessment worksheets. Consistency, consistency impact, accuracy, and accuracy impact were measured using the formulas presented in Chapter IV.C. Assessment. The following is the null hypothesis.

H_0 : SMEs demonstrate the same levels of effect on consistency and accuracy during validation of an HBR model implementation using a 7-Point Likert Scale as they do when using a 5-Point Likert Scale or Go/No-Go Scale.⁵⁵

Conversely, the alternative hypothesis is:

H_A : At least one scale (7-Point Likert, 5-Point Likert, or Go/No-Go) produces different effects on SME consistency and accuracy during validation of an HBR model implementation.

Considering these individually, the null hypothesis is

$$H_0 : \mu_{I_7} = \mu_{I_5} = \mu_{I_G}$$

versus the alternative hypothesis

$$H_A : \text{at least one } \mu_{I_i} \neq \mu_{I_j}$$

where

H_0 : Null Hypothesis,

H_A : Alternative Hypothesis,

μ_{I_7} : The average level of the issue (inter-SME consistency, intra-SME consistency, intra-SME consistency impact, intra-SME accuracy, and intra-SME accuracy impact) demonstrated by SMEs assessing performance using a 7-Point Likert Scale,

⁵⁵ The modified scales are a traditional Go/No-Go used in ARTEP evaluations worksheets (Appendix K. Assessment Worksheets) and a 5-Point Likert Scale.

μ_{I_s} : The average level of the issue (inter-SME consistency, intra-SME consistency, intra-SME consistency impact, intra-SME accuracy, and intra-SME accuracy impact) demonstrated by SMEs assessing performance using a 5-Point Likert Scale,

μ_{I_g} : The average level of the issue (inter-SME consistency, intra-SME consistency, intra-SME consistency impact, intra-SME accuracy, and intra-SME accuracy impact) demonstrated by SMEs assessing performance using a Go/No-Go Scale,

μ_{I_i} : The average level of the issue (inter-SME consistency, intra-SME consistency, intra-SME consistency impact, intra-SME accuracy, and intra-SME accuracy impact) demonstrated by SMEs assessing performance using a 7-Point Likert Scale, a 5-Point Likert Scale, or a Go/No-Go Scale, and

μ_{I_j} : The average level of the issue (inter-SME consistency, intra-SME consistency, intra-SME consistency impact, intra-SME accuracy, and intra-SME accuracy impact) demonstrated by SMEs assessing performance using a 7-Point Likert Scale, a 5-Point Likert Scale, or a Go/No-Go Scale and different from the scale used to calculate μ_{I_i} .

2. Design

There were five dependant variables for these studies: inter-SME consistency, intra-SME consistency, intra-SME consistency impact, intra-SME accuracy, and intra-SME accuracy impact.⁵⁶ The studies seek to determine if inconsistency and inaccuracy demonstrated by SMEs identified in the assessment of perceived human performance and if perceived CGFs controlled by HBR models can be mitigated using varied assessment scales independent of SME personality, cognitive model, and scenario.

These studies had the same primary and secondary factors as first study with the addition of the primary factor, scale. The three scales used were a 7-Point Likert Scale,

⁵⁶ Section IV.C. Assessment describes the experiment formulas for inter-SME consistency, intra-SME consistency, intra-SME consistency impact, intra-SME accuracy, and intra-SME accuracy impact.

5-Point Likert Scale, and Go/No-Go Scale. Table 8 denotes the type and levels of primary factors. Of seven factor types, three had more than one level: behavior generator, validation methodology by scale, and scenario.

Table 8. Study #2, Primary Factor Levels

Reference ID	Factor	Number of Levels
A	Behavior Generator	2
B	Validation Methodology (Scale)	3
C	Subject Matter Expert	1
D	Model	1
E	Domain	1
F	Scenario	3
G	Simulation/ Game	1

The factors for this study are behavior generator belief, assessment scale, and scenario (Table 9). Table 9 indicates the experimental layout for second study by SME groups and treatments. The treatments for simulation belief were perceived human and simulated performance. The treatments for scenario were two offensive and one defensive scenarios. The levels for scale were 7-Point Likert, 5-Point Likert, and Go/No-Go scales. The SME Groups represent the study blocks.

Table 9. Experimental Layout for Study #2

SME Group	Offensive Scenario #1	Defensive Scenario	Offensive Scenario #2
1	7 Pt Simulated Performance	7 Pt Simulated Performance	7 Pt Simulated Performance
2	7 Pt <i>Human Performance</i>	7 Pt <i>Human Performance</i>	7 Pt <i>Human Performance</i>
3	Go-No Go Simulated Performance	Go-No Go Simulated Performance	Go-No Go Simulated Performance
4	<i>Go-No Go Human Performance</i>	<i>Go-No Go Human Performance</i>	<i>Go-No Go Human Performance</i>
5	5 Pt <i>Human Performance</i>	5 Pt <i>Human Performance</i>	5 Pt <i>Human Performance</i>
6	5 Pt Simulated Performance	5 Pt Simulated Performance	5 Pt Simulated Performance

3. Procedures, Set-Up, and Study Phases

Procedures used were the same as those for the first study. The only differences were the lack of a pilot study, the addition of research assistants for data collection, and the location. Data collection was conducted from 18 September 2003 through 24 September 2003 and from 22 through 29 October 2003 at Fort Benning, GA. Respectively, the locations for data collection were Room O-256 and Class Room 52, ICCC, Building #4, Fort Benning, GA. The second facility differed from the first in size and number of SME workstations. The additional October study was conducted in a 50 foot by 50 foot room accommodating 27 tables large enough for individual SME workstations (Figure 20). Study material and equipment were arranged similarly to the first study. Figure 21 is a picture of the room and its setup prior to the arrival of SMEs. The image was taken from the front row of tables.

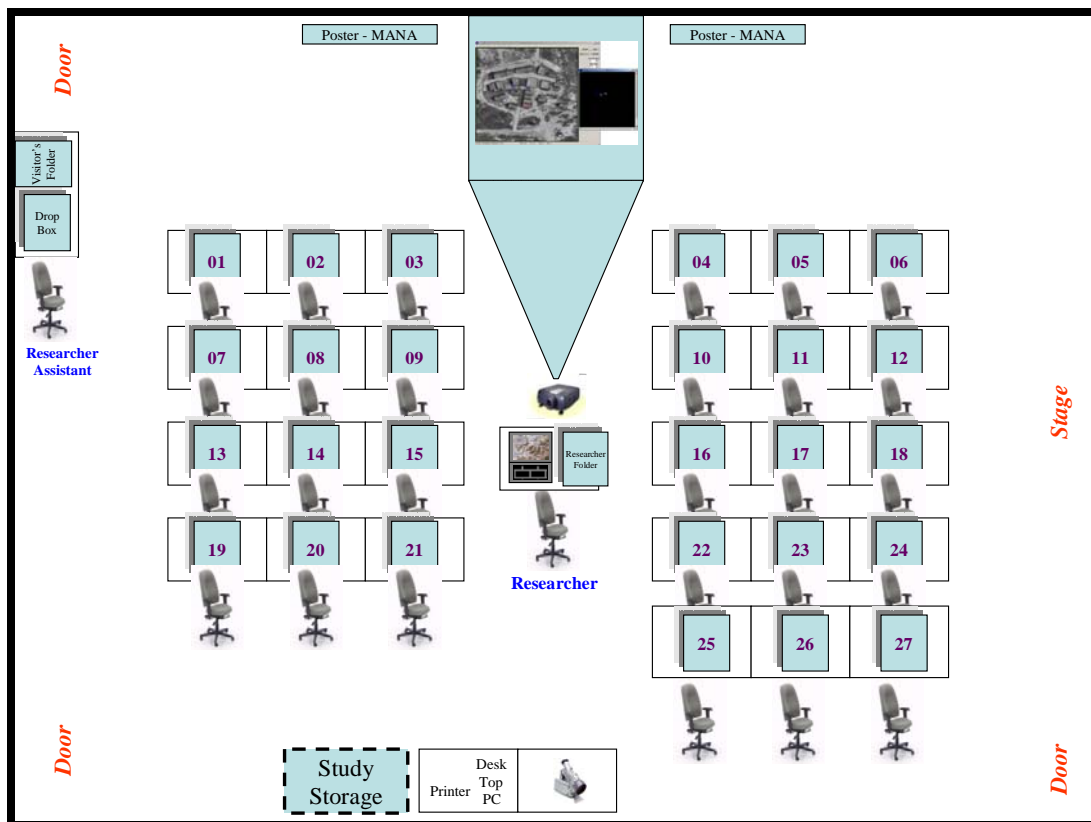


Figure 20. Room Layout for October Data Collection



Figure 21. Room Setup for October Data Collection

As with the first study, these studies were conducted in five phases: in-processing, familiarization, training, data collection, and debrief. The only difference between the studies was the scales used for the assessment worksheets.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. RESULTS

This chapter addresses analysis of data collected during the studies involved in this research and is divided as follows. Section A describes the SMEs, the source of assessment data, using SME demographics and NEO-FFI data. Next, Section B addresses the presence and effects of bias. The next section presents the results of the face validation process and investigates significant factors in the data set in terms of the study hypothesis. Finally, knowing bias, scale, and simulation belief are not the only factors potentially influencing the results of face validation, the chapter looks at some demographic factors that affect the consistency and accuracy of results.

The statistical package used to analyze the data for this research is JMP®; The Statistical Discovery Software by SAS Institute, Inc. Data collected from SMEs is entered into Excel® spreadsheets before being imported into JMP® for analysis. The ordinal (categorical) responses allow analysis based on nonnormal distributions vice the normal distribution; the data for this study is categorical. Analysis of categorical data (ANOCAT) is employed to identify factors that are statistically significant with respect to defined responses. The factors examined are bias, assessment scale, and simulation belief. Stepwise logistical regression is used to identify additional factors that appear to affect SME accuracy impact, meaning they effect the overall assessment of the model.

A. GENERAL

This section includes a discussion of SME demographics, content, and organization of the datasets. This research generalizes data from two pilot studies and two base studies. The studies involved 5 pilot study SMEs and 182 base study SMEs.

1. Subject Matter Expert Demographics

The sample for this research consists of 182 SMEs from the population of two ICCC classes. The SMEs were all male. Table 10 presents SME time in service data broken down by the assessment scale utilized. SMEs ranged from 25 to 41 years of age with a mean age of 28.88 years. SME time in service ranged from three to 20 years with a mean of 7.23 years of service. Over 41% of the 74 SMEs had served as enlisted personnel. The mean enlisted time served by those with prior service was nearly five

years. The time since these officers were last with troops ranged from zero months to seven years with a mean of ten months and a median of three months.

Table 10. Subject Matter Expert Demographics: Time in Service Data⁵⁷

Statistic	Assessment Scale			
	All Scales	7-Point	Go/No-Go	5-Point
Total Number of SMEs using Assessment Scale	182	80	50	52
Age (Years)				
Number of Respondents	181	80	50	51
Min Response (Years)	25	25	26	25
Max Response (Years)	41	41	38	37
Mean Response (Years)	28.88	29.14	28.9	28.47
Total Time of Service (Months)				
Number of Respondents	178	78	49	51
Min Response (Months)	36	38	40	36
Max Response (Months)	240	237	240	194
Mean Response (Months)	87.78	92.87	83.12	84.47
Total Time in Service (Years)				
Number of Respondents	178	78	49	51
Min Response (Years)	3	3	3	3
Max Response (Years)	20	20	20	16
Mean Response (Years)	7.23	7.68	6.84	6.92
Prior Enlistment Time (Months)				
Number of Respondents	178	77	49	51
Min Response (Months)	0	0	0	0
Max Response (Months)	194	194	144	150
Mean Response (Months)	29.99	33.55	27.67	26.86
Time Since in Line Unit (Months)				
Number of Respondents	178	78	49	51
Min Response (Months)	0	2	0	0
Max Response (Months)	84	84	42	41
Mean Response (Months)	10.24	12.96	7.51	8.71

Table N.1 in Appendix N. Data Analysis shows the military service and level of education data for the SMEs. As the table indicates, 178 were US Army officers while four were US Marine Corps officers. As officers, all SMEs had earned bachelors degrees with six possessing master's degrees. One-hundred and seventy-four held the rank of

⁵⁷ Data excludes participants who did not respond to the specific question(s) on the Participant Demographics Questionnaire.

Captain while six were senior First Lieutenants. A majority of the officers, 159, were primarily Infantry officers. The Army had designated twenty-one of all the officers for transfer to the Special Forces. SMEs have served in all of the traditional infantry units with 48 (26.37%) serving in more than one type of unit such as a light infantry division and the Ranger Regiment. Thirty-nine had served in light infantry units, 51 in airborne units, 38 in the 101st Air Assault Division, 58 in mechanized units, six in the Sticker Brigade, and 23 in the Ranger Regiment or Special Forces. Thirty-one had served in other types of units to include armor, artillery, or corps level support units. One hundred and forty-nine SMEs (81.87%) had held the position of rifle platoon leader, a position responsible for training squads and evaluating fire teams. SMEs had held numerous other duty positions from individual rifleman to company commander. Table N.1. lists these positions.⁵⁸ Ninety-nine had held 38 other positions not specifically listed in Table N.1.

SMEs had varied real-world experience as shown in Table 11. The percentages shown in Table 11 are based on the total number of SMEs and number of SMEs per specific scale group. Sixty-nine percent had deployed with their units to execute combat or peacekeeping missions. Fifty-one percent had been in combat in Panama, Afghanistan, and/or Iraq. Of the 93 SMEs who had been in combat, 83 of them had been in units that executed urban operations under combat conditions.

Table 11. Subject Matter Expert Demographics: Deployment Data

	Total (182)		7-Point (80)		Go/No-Go (50)		5-Point (52)	
	(#)	(%)	(#)	(%)	(#)	(%)	(#)	(%)
Peacekeeping or Combat Experience	126	69.23%	48	60.00%	33	66.00%	32	61.54%
Combat Experience	93	51.10%	39	48.75%	33	66.00%	29	55.77%
Urban Combat Experience	83	45.60%	39	48.75%	27	54.00%	22	42.31%

Subject matter expert experience with video games and combat models was limited as shown in Table 12. Nearly 70% of the SMEs characterized themselves as having played only one to ten hours of video games a month over the past two years. At the time of the research, 73% spent no time playing video games. In fact, 65% reported they had no experience or consider themselves novices with first shooter video games.

⁵⁸ Participants can hold more than one duty position.

Combat model experience is even more limited. Although 102 SMEs report being exposed to at least one combat simulation, most had less than two days exposure to combat simulations. Of the 62 SMEs who expressed an opinion on the human behavior representation of combat models, 4.8% gave a positive assessment, 37.1% an average appraisal, 43.2% a negative, and 16.1% had no impression; only one SME (1.6%) gave a mixed assessment.

Table 12. Subject Matter Expert Demographics: Game and Model Experience

Statistic	Assessment Scale			
	All Scales	7-Point	Go/No-Go	5-Point
Total Number (N) of SMEs using Assessment Scale	182	80	50	52
Combat Models Experience				
Yes	102	44	32	26
No	71	29	18	24
Failed to Respond	9	7	0	2
Combat Models Days of Use				
Number (N) of Responses	77	44	26	17
Min (Days)	1	1	1	1
Max (Days)	63	63	40	20
Mean (Days)	6.99	6.68	8.46	5.35
Video Game Experience (2 years)⁵⁹				
Number (N) of Responses	175	74	50	51
Min (Hours/Year)	0	0	0	0
Max (Hours/Year)	208+	208+	208+	208+
Median (Hours/Year)	1-10	1-10	1-10	1-10
Video Game Experience (Current)⁶⁰				
Number (N) of Responses	175	74	50	51
Min (Hours/Week)	0	0	0	0
Max (Hours/Week)	4+	2-3	4+	4+
Median (Hours/Week)	0	0	0	0

The NEO-FFI was a means to codify SME personalities. The purpose of codifying the SME personalities is to provide a starting point for gauging personality effects on SME accuracy impact when appraising perceived human or simulation

⁵⁹ Video Game Experience (2 years) refers to the number of hours spent playing video games over the past two years. The responses were broken down into five bins: 1: none, 2: 1-10 hrs a year, 3: 1-10 hrs a month, 4: 3-4 hrs a week, and 5: 4+ hrs a week.

⁶⁰ Video Game Experience (Current) refers to the current number of hours spent each week playing video games. The responses were broken down into six bins: 1: none, 2: 0.5-1 hrs per week, 3: 1-2 hrs per week, 4: 2-3 hrs per week, 5: 3-4 hrs per week, and 6: 4+ hrs per week.

performance. These personality differences accounted, in part, for the differences observed in individual responses where some individuals' assessment scores are more accurate than others.

Mean SME scores were not identically distributed with respect to the general male public; within the SME sample, the conscientiousness, openness, and agreeableness scores are in line with the general US male population. The mean score trends for neuroticism, extraversion, and conscientiousness are similar to those reported in the Buziak, 2000; Wellbrink, 2003; and Miller & Shattuck, 2004 studies where participants were also military personnel. Miller & Shattuck's participants reported a lower mean score for openness (less open) than the participants in the Buziak, Wellbrink, and this research. Buziak's participants reported a higher agreeableness mean score (more agreeable) than participants in the Wellbrink, Miller & Shattuck, and this research. These disparities may be due to differences in the number of participants and their military services. Buziak's 116 participants were US Naval officers, Miller & Shattuck's eight participants were US Army personnel (Majors, Captains, and mid level Non-Commissioned Officers), Wellbrink used a mix of fifty male and female, US and International officers from three different military services, and in this research, 178 (97.8%) of the SMEs were US Army company grade officers [BUZI 00] [WELL 03] [MILL 04].

Appendix N., Table N.3 shows the NEO-FFI mean raw scores for all SMEs who completed the inventory for this research. The lower mean raw value of neuroticism, 11.93 (42%), indicates experiment SMEs were more relaxed and secure with themselves than the average US male who scored 17.60 (50%). The higher raw score mean extraversion value, 32.00 (58%) versus 27.22 (50%), suggests SMEs were more sociable, friendly, and talkative than the average US male. The mean openness raw score for SMEs and the average US male were effectively the same, 27.27 (50%) and 27.09 (50%) respectively. The slightly lower mean agreeableness raw score, 30.13 (46%) for SMEs and 31.93 (50%) for the average US male, indicates SMEs were slightly more critical and potentially more confrontational; however, the SMEs' mean scores were not statistically different based on a 2-sample t-test where the confidence intervals include both values.

The SMEs' slightly higher mean raw score of 35.28 (52%) for conscientiousness was still in the average range but suggests the average SME was more organized and careful than the average US male, 34.10 (50%). In summary, the average SME was more controlled, amiable, thoughtful, organized, and observant than the average US male.

2. Data Set

Table 13 shows the breakdown of the 182 SMEs into six simulation belief and scale groups (Sim-Scale Group). The first digit (0 or 1) of the Sim-Scale Group indicates the SME's simulation belief. Simulation belief *0* represents those SMEs who were told a computer model was generating the behaviors they were observing, i.e., constructive simulation. Simulation belief *1* represents those SMEs who were told soldiers, affixed with tracking instrumentation, performed the tasks at the McKenna MOUT Site to generate the behaviors, i.e. live simulation. The second digit (1, 2, or 3) is the assessment scale utilized by the SME: 7-Point Likert Scale, Go/No-Go Scale, or 5-Point Likert Scale, respectively. All scales were applied to the MTP-based assessment forms.

Table 13. Data Set Groupings

Sim-Scale Group	Simulation	Scale	# of Sample SMEs
Sim-Scale 0-1	0	1	40
Sim-Scale 0-2	0	2	25
Sim-Scale 0-3	0	3	25
Sim-Scale 1-1	1	1	40
Sim-Scale 1-2	1	2	25
Sim-Scale 1-3	1	3	27

SMEs assessed the performance of the human or simulated human performance on four levels. The overall assessment was based on the assessment of three scenarios: two offensive and one defensive. Each scenario had one to three tasks to be assessed (seven overall tasks across the three scenarios with three distinct tasks). Finally, each task included 15 - 21 possible subtasks to be assessed. Figure 22 depicts the relationship between the different levels of assessment. Assessment of the behaviors starts with subtasks, which aggregate into tasks, tasks aggregate into scenarios, and the scenario assessments amass in the overall assessment. The terms sublevel and level responses are used when speaking of general concepts or formulas that apply to more than one pairing.

A sublevel is the stage directly below the current level. For example, subtasks are the sublevel of the level task, and tasks are the sublevel of the level scenario.

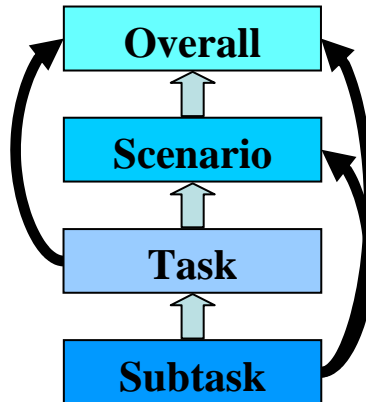


Figure 22. Experiment Assessment Sublevels and Levels

Data analysis for this research focuses on SME responses for each subtask, task, scenario, and overall assessment question. Each subsequent section deals with a specific characteristic of the data. The sections addressing consistency and consistency impact deal with the level and sublevel pairs displayed in Figure 22: subtask to task, task to scenario, scenario to overall, subtask to scenario, subtask to overall, task to overall. The accuracy and accuracy impact sections deal with SME scores at each level of assessment: subtask, task, scenario, and overall.

B. BIAS PATTERNS

As described in Section II.F.2. Bias, bias is systematic error introduced into the rating process by a SME who consistently selects one response over another, disregarding the actual information presented to him. Bias manifests itself in assessment in three ways: increasing effects, decreasing effects, or modulating effects. Figure 23 illustrates these three effects on the consistency or accuracy impact. It demonstrates how the effects of bias may have compounding or mitigating impact on the final assessment.

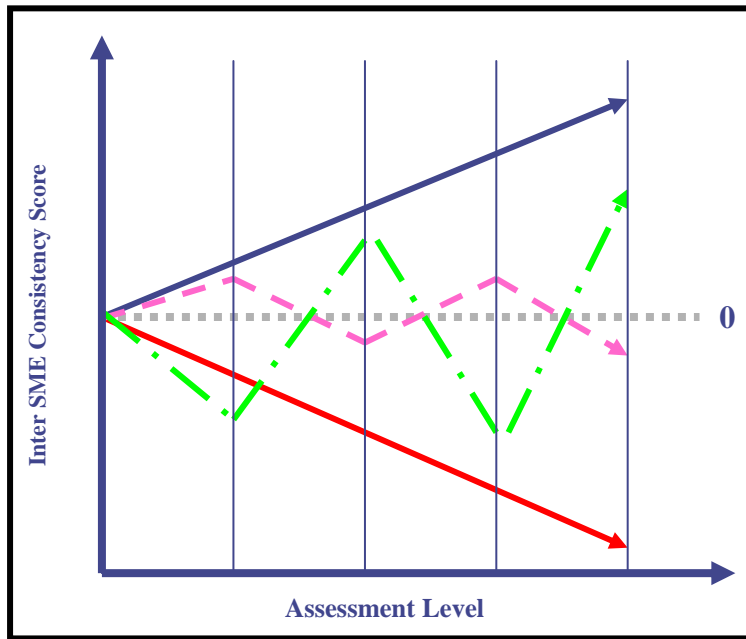


Figure 23. Bias Manifestations

SMEs demonstrate *performance bias* for two reasons. First, a SME was unable to make assessments due to the availability of data. Second, the SME lacked the ability or desire to comply with specified validation procedures. These two reasons manifest themselves in SME responses. A SME who chose not to provide definitive responses to 20% or more of the assessment questions is categorized as displaying performance bias.⁶¹

Figure 24 illustrates an example performance bias response pattern. Of a possible 159 questions, SME B2124 provided only 16 (10.06%) responses that were not “Not Applicable” or “No Opinion.” Based on his comments, B2124 felt the simulation failed to furnish enough information for him to make an assessment. SME B2223’s only responses were explanations of his aversion to the simulation interface and the unrealistic nature of the scenarios. Along with his singular response value to all NEO-FFI personality test questions, B2223 demonstrated a lack of desire to participate in the validation process.

⁶¹ A definitive response is a “Go” or “No-Go” assessment of the subtask, task, scenario, or overall assessment question. “Not Applicable” or “No Opinion” responses are not definitive responses.

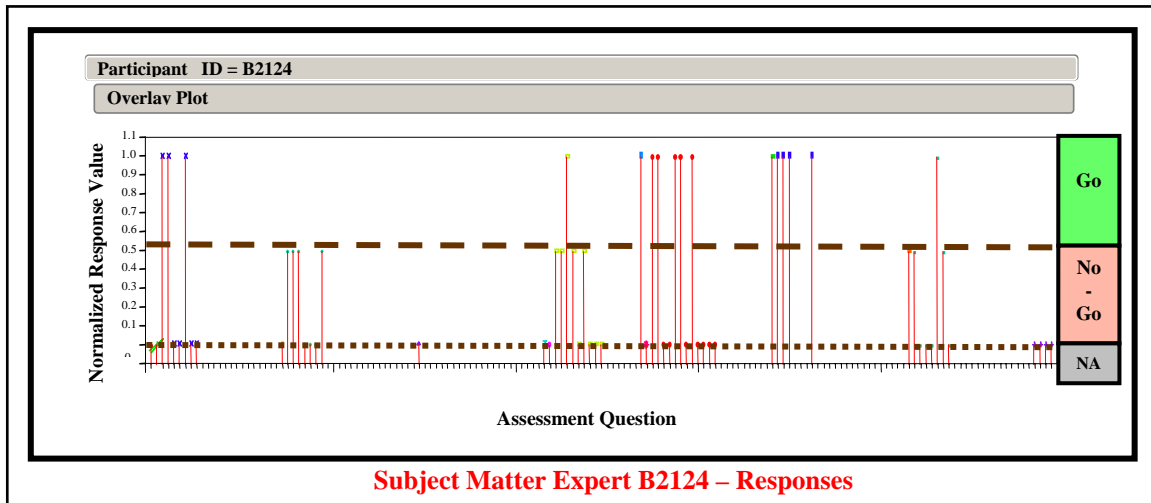


Figure 24. Performance Bias Example

Of the 182 SMEs, 23 (12.63%) displayed performance bias. Excluding these SMEs from the sample data results in the same or higher mean scores for the overall assessment questions across all scales, simulation beliefs, and scale-simulation belief groupings (Appendix N., Table N.12). Thus, although Sim-Scale Groups with mean sample scores categorized as “No-Go” were less consistent, global inter-SME results were more consistent. Consistency here indicates that normalized mean scores assessed as “Go” in the original sample settings had higher normalized mean assessment scores when SMEs identified as displaying performance bias are excluded from the analysis. Conversely, when those SMEs displaying performance bias were excluded from the sample those normalized overall mean scores assessed as “No-Go” in the original sample settings had lower normalized mean scores and thus were more consistent.

In this data, *anchoring bias* was identified in a SME’s raw data in one of two ways. First is when a SME judges the first task, and associated subtasks, as a “Go”, and then after viewing the second task and associated subtasks, which were not performed correctly, judges the remainder of the model performance as “Go” for more than 90% of the assessment questions.⁶² Second is when a SME judges the first scenario, associated tasks and subtasks, as “No-Go”, and then after viewing the second scenario and

⁶² In accordance with doctrine, the squad failed to perform properly the second task, and associated subtasks for “React to the Sniper Attack,” by losing two personnel without the remainder of the squad reacting to the sniper’s attack or the loss of personnel.

associated subtasks judges the remainder of the model performance as “No-Go” for more than 90% of the assessment questions for which he provides a passing or failing appraisal.⁶³

Figure 25 provides illustrations of two different response patterns that are examples of anchoring bias. SME B1107 starts out by assessing the first task and associated subtasks as “Go”. After the second task and associated subtasks, which he assessed as a “No-Go”, he provides zero “No-Go” responses for the remainder of the assessment. SME B2204 shows the exact opposite effect providing only eight “Go” responses for the entire assessment, with only three “Go” responses after viewing and assessing the second scenario.

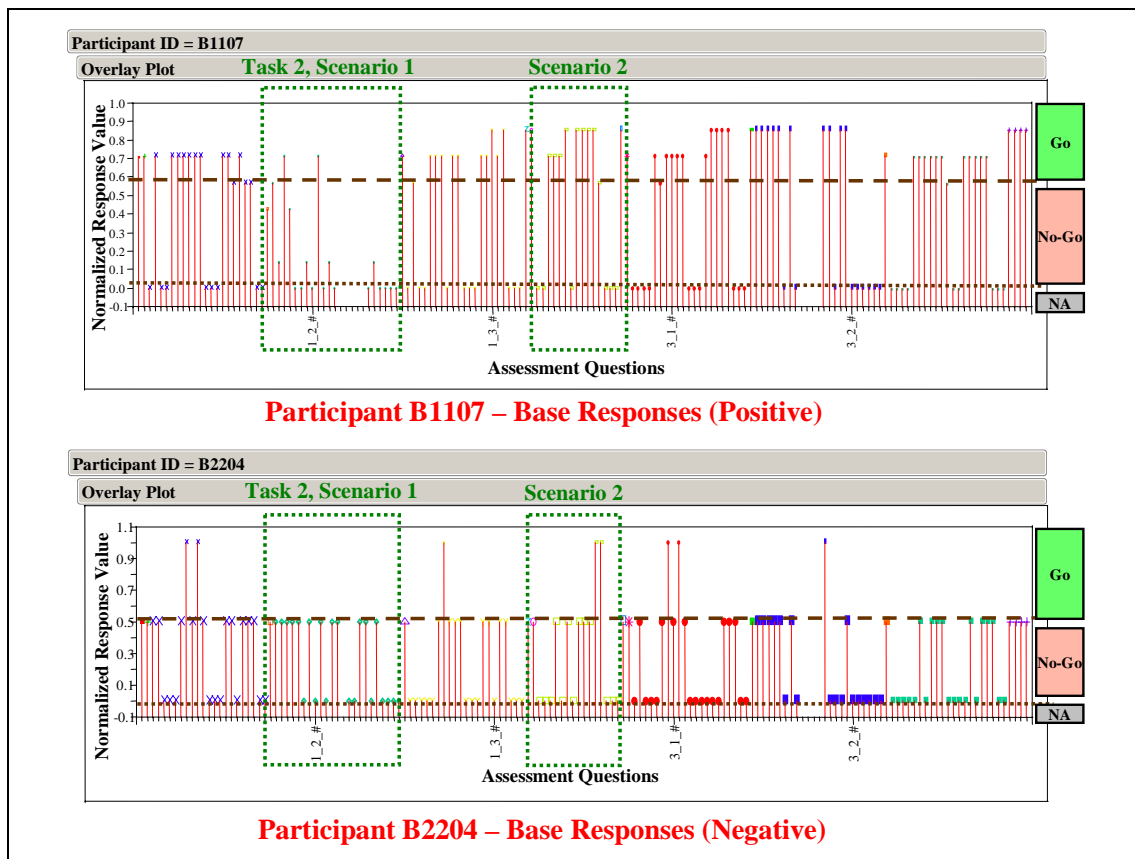


Figure 25. Anchoring Bias Examples

⁶³ In accordance with doctrine, the squad properly performed the second scenario and associated task and subtasks as it successfully defends the building by destroying enemy forces attempting to seize the structure.

Thirty SMEs (16.48%) displayed anchoring bias. Nineteen of the 24 Sim-Scale Group mean scores for overall assessment questions increase in value. Of those with decreasing means, only one (Sim-Scale 1-1, Overall Question 2) is not more consistent.⁶⁴ Across the four general mean overall assessment scores, mean values either increased or are more consistent when SMEs with anchoring bias are excluded from the sample data (Appendix N., Table N.13).

Contrast bias is identified after inspecting two formats for SME data. The first data pattern is found in a plot of a SME's raw data. Potential contrast bias was identified when a SME started with either a negative or positive opinion and after viewing data, which differs from this initial opinion, the SME negates any further evidence in support of the original hypothesis and assesses the model based on the swing of opinion. The second data pattern is found in a plot of a SME's accuracy scores. The plot of the accuracy data indicates if a SME shifts in his accuracy trend, from harsher to more lenient or from more lenient to harsher, as the assessment process proceeds. To demonstrate contrast bias, this shift occurs after the swing in raw score responses.

Figure 26 is a combination of a SME's raw data and accuracy plots that demonstrate patterns, which together exemplify contrast bias. SME B1109 starts by assessing the first task and associated subtasks as "Go". After the second task and associated subtasks, which he assessed as a "No-Go", he provides a higher percentage of "No-Go" responses for the remainder of the assessment. The SME's accuracy score plot illustrates that nine of the first 45 responses (20.00%) were harsher than the key assessment responses. However, after assessing task number two, the SME scored 65 of the remaining 114 responses (57.02%) harsher.

⁶⁴ Consistency is defined in the same manner as for performance bias.



Figure 26. Contrast Bias Example

Five SMEs (2.75%) displayed contrast bias. Changes in overall assessment scores are identified when SMEs with contrast bias are excluded from the sample data except for the constructive simulation belief, 5-Point Likert Scale group (Sim-Scale 0-3), which had no SMEs identified with contrast bias. Within those Sim-Scale Groups that did change, all mean scores for the overall assessment questions are higher in value or more consistent for all scales, simulation beliefs, and scale-simulation belief groupings (Appendix N., Table N.14).⁶⁵

Confirmation bias manifests itself in a plot of a SME's consistency scores per sublevel-level pairing. When a SME feels certain factors are more important than other factors, a result can be a majority of responses indicating one overall assessment but the SME making a different final assessment, the whole is not equal to the sum of the parts. Confirmation bias manifests itself in the data in two forms. First is when differences

⁶⁵ Consistency is defined in the same manner as for performance bias.

between sublevel mean scores and level responses tend toward no difference in response but the overall response differs.⁶⁶ Second is when differences between sublevel mean scores and level responses show a general trend in being harsher or more lenient but the overall response differs from this trend. Inconsistent results could identify the task within which lies the factor or factors that are providing an excessive amount of influence on the overall assessment.

Figure 27 illustrates two different response patterns that are examples of confirmation bias. SME B1311 shows either consistent or positive inconsistency in his sublevel-level pairing results for all but one on the pairings below the scenarios-overall assessment pairings. This indicates one or more of the subtasks in the second scenario overly influenced the SME's overall model appraisal and is interpreted as confirmation bias. SME B1316 is either consistent or harsher in his sublevel-level pairing results for all but one on the pairings below the scenarios-overall assessment pairings. This indicates one or more of the subtasks in the third scenario overly influenced the SME's overall assessment of the model and is interpreted as confirmation bias.

⁶⁶ Note: differences between sublevel (sublevel, task, or scenario) mean scores and level (task, scenario, or overall, respectively) responses may be mitigate by additional assessment responses; this does not indicate confirmation bias.

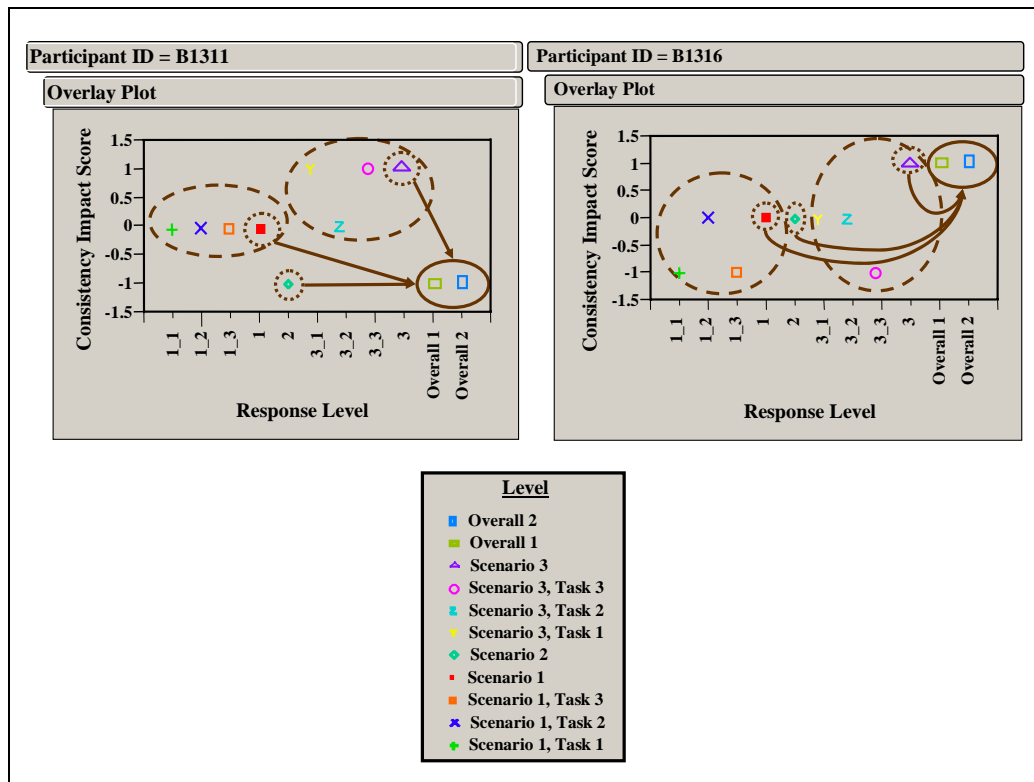


Figure 27. Confirmation Bias Examples

Data from 55 SMEs (30.22%) is interpreted as displaying confirmation bias. Except for results where groups were using the 5-Point Likert Scale, all mean scores for the overall assessment questions increased in value. However, 35 (79.55%) of the group, overall response, mean scores are more consistent when SMEs with confirmation bias are excluded from the sample data (Appendix N., Table N.15).⁶⁷ For those three groups using the 5-Point Likert Scale, all but Sim-Scale 1-1 is more consistent.

Figure 28 displays the results of bias identified amongst SME responses from the initial study. SMEs using the 7-Point Likert Scale demonstrated the same number of bias cases whether they believed they were assessing simulated behaviors or human behaviors.

⁶⁷ Consistency is defined in the same manner as for performance bias.

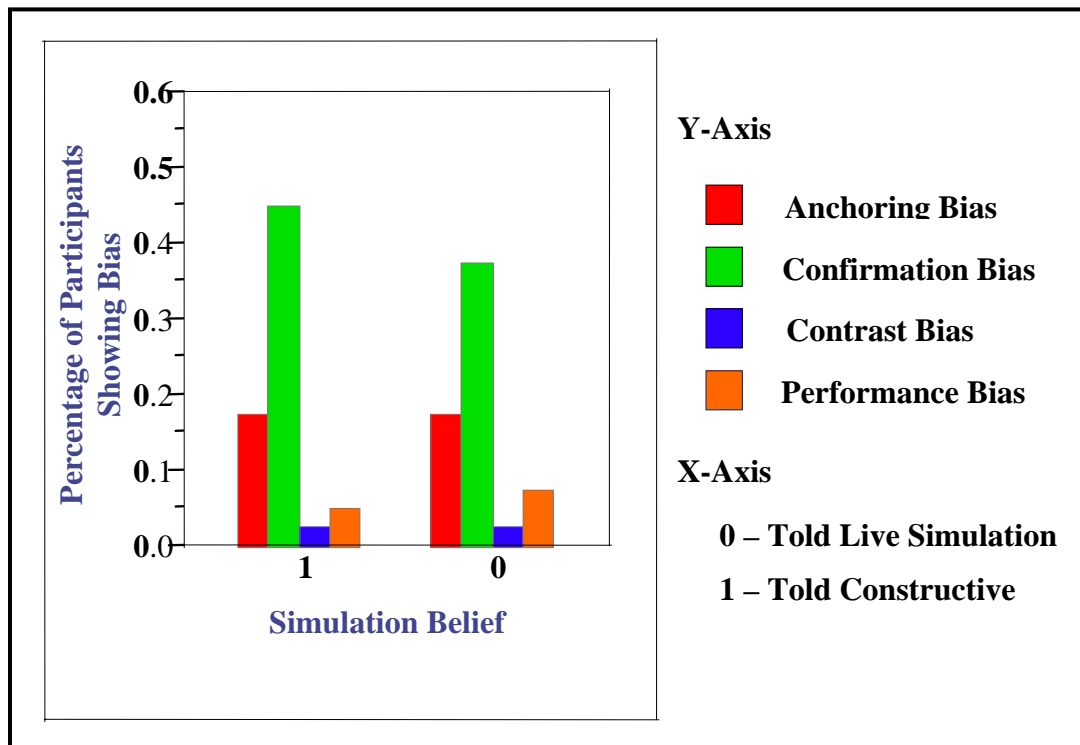


Figure 28. Study #1, Subject Matter Expert Bias for 7-Point Likert Scale

C. ASSESSMENT

The *overall assessment* is the result of SME raw scores for each of the four overall assessment questions. This result is measured by calculating the mean score for the normalized (0 to 1) SME responses for each question. Normalized mean scores equal to, or greater than, 0.667 are categorized as “Gos” or valid behaviors. Values above 0.667 are chosen since all normalized values above this score fall into the range of responses which SMEs are told are passing scores. Overall 1 is the SMEs’ assessment of the performance of individual soldier skills. Overall 2 is the SMEs’ assessment of the squad leaders’ performance. Overall 3 and Overall 4 are predictive assessments of the quality or realism of the behaviors as SMEs assess the individual soldier skills and squad leaders’ performance as if the tasks are to be performed in an environment not previously assessed (e.g. artic, desert, jungle, etc.).

Table 14 displays the overall assessment results for the performance of the model based on group mean scores. This table indicates that, when considering both

performance assessment scores (Overall 1 and Overall 2), three of the six groups rated the model as being valid for the tasks and scenario assessed. For the predictive overall assessment scores (Overall 3 and Overall 4), only the two groups using the Go/No-Go Scale rated the model or human behaviors as valid. Of the overall assessment scores not receiving a valid score, only the live simulation belief (0) and 5-Point Likert Scale (3) group rated the model as invalid, scores less than 0.500. Scores less than 0.500 are the normalized scores that fall into the range of responses which SMEs are told are failing scores.

Table 14. Mean Values for Normalized, Overall Assessment Scores

ID		Number of SMEs		Mean (Normalized 0-1 Responses)			
Simulation Belief	Scale	Overall 1 & Overall 2	Overall 3 & Overall 4	Overall 1	Overall 2	Overall 3	Overall 4
0	1	37	36	0.583	0.598	0.540	0.552
0	2	25	25	0.920	0.920	0.920	0.940
0	3	24	24	0.483	0.500	0.442	0.433
1	1	39	39	0.667	0.696	0.593	0.623
1	2	25	25	0.820	0.820	0.780	0.800
1	3	25	25	0.616	0.664	0.600	0.632
All Beliefs and Scales		175	174	0.675	0.694	0.636	0.654

The raw scores, although not normally distributed, have standard deviations ranging from 0.1830 to 0.3559 for the overall assessment scores indicating a high degree of variability. Figure 29 is an example of the variability in distribution of SME responses for the same realization to observe. The x-axis is the normalized raw response score and the y-axis displays the number of SMEs who assessed the question, Overall 1, at the score value. Scores range between 0 and 1 with the mean equal to 0.675. The median (0.714) falls to the left of the mean and the responses are not normally distributed. Although nearly 60% of the scores indicate a “Go” status, 26% of the scores fall into the “No-Go” category with the remaining 14% of the scores being “Not Applicable” or “No Opinion” responses.

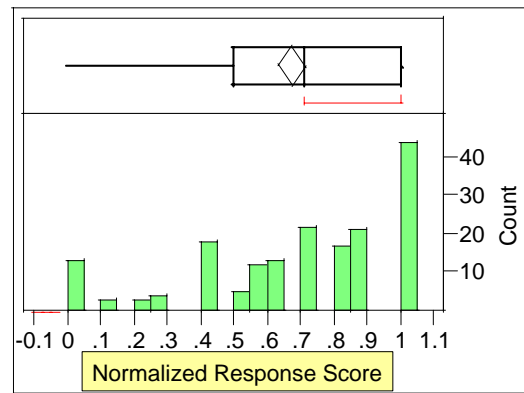


Figure 29. Distribution of Subject Matter Experts' Normalized Responses to Question Overall 1

The degree of SME variance depicted in Table 14 indicates there is an issue with inter-SME consistency. Inter-SME consistency refers to the agreement between SMEs when they rated each subtask, task, scenario, and overall question rating. This inconsistency is identified by examining the variability in SME responses for each question. Figure 30 and Figure 31 illustrate inter-SME consistency between SME responses when observing and assessing the same behavior event via the model interface. This inconsistency precludes the likelihood of an accurate assessment of the simulation. Extensive analysis of plots for each assessment question was performed, but only two are shown here. Fifty (31.45%) subtasks, tasks, scenarios, and overall assessment responses plots exhibit inconsistent responses distributions similar to Figure 30 and Figure 31.

Figure 30 and Figure 31 display the effect of scale and simulation belief on the SME responses when observing and assessing the same behavior event through the model interface. The figures display response by scale and simulation belief based on shape and color code. Red “□” are assessment scores for SMEs who believe they are observing constructive behaviors and are using the 7-Point Likert Scale. Red “■” are assessment scores for SMEs who believe they are observing recorded live behaviors and are using the 7-Point Likert Scale. Green “◇” are assessment scores for SMEs who believe they are observing constructive behaviors and are using the Go/No-Go Scale. Green “+” are assessment scores for SMEs who believe they are observing recorded live behaviors and

are using the Go/No-Go Scale. Blue “△” are assessment scores for SMEs who believe they are observing constructive behaviors and are using the 5-Point Likert Scale. Blue “X” are assessment scores for SMEs who believe they are observing recorded live behaviors and are using the 5-Point Likert Scale.

Figure 30 illustrates no observable difference in the distribution of assessment scores based on scale or simulation belief. Figure 31 illustrates no observable difference in the distribution of assessment scores based on simulation belief. However, Figure 31 does illustrate an observable difference in the distribution of assessment scores based on scale. Those using the Go/No-Go Scale, “◇” and “+”, provide a higher percentage of valid assessment scores than the other two scales. Those using the 5-Point Likert Scale, “△” and “X”, provide a higher percentage of invalid or “Not Applicable” assessment scores than the other two scales. These results are similar to those seen by the mean scores in Table 15

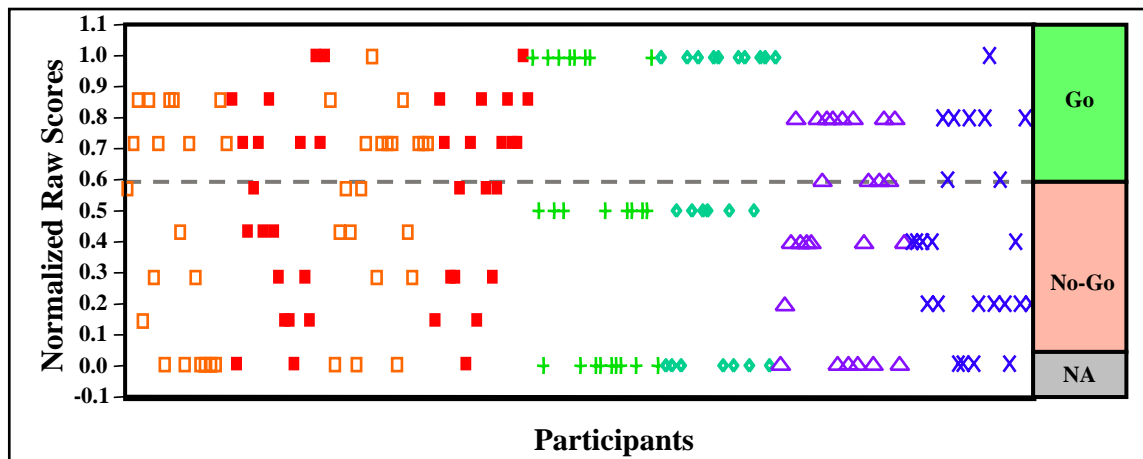


Figure 30. Subject Matter Expert Normalized Responses to Subtask 2, Task 1, Scenario 1

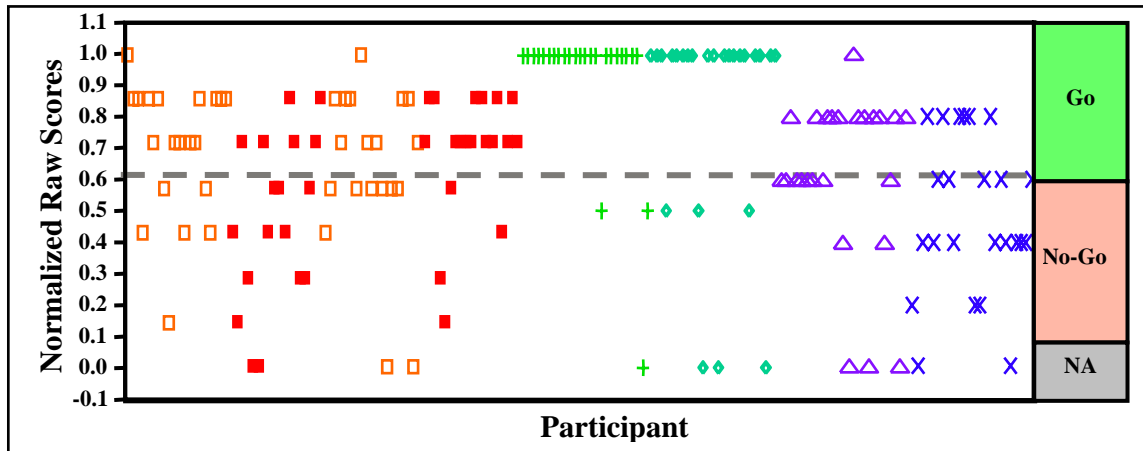


Figure 31. Subject Matter Expert Normalized Responses to Overall 1

Four separate ANOCATs are performed for each assessment level: subtask, task, scenario, and overall. In each case, the response was the normalized assessment rating averaged across level. Factors considered are scale and simulation belief, where scale equals the assessment scale used by the SMEs and simulation belief equals whether the SMEs are told the process they are observing is based on live or simulated performance. Recall, scale has three levels: 7-Point Likert, Go/No-Go, and 5-Point Likert. The model employed for analysis considered the main effects of the two factors, scale and simulation belief, and an interaction effect (scale cross simulation belief). Table 15 shows the results of the four analyses. For the analysis, $\alpha = 0.05$. Thus values of Prob>ChiSq less than 0.05 indicate the factor is statistically significant.

Data indicates one or more factors is statistically significant at each level of assessment with the Whole Model Test Prob>ChiSq equal to or less than 0.0001 in each case. Looking at scale as a factor, the table indicates a statistically significant effect for all levels with the Effect Likelihood Ratio Test's Prob>ChiSq equal to 0.0000. For simulation belief, no statistically significant effect is present. Finally, there is a statistically significant effect based on scale cross simulation belief (Prob>ChiSq is less than 0.05) at every level. This indicates interactions are present between the two factors. However, in examining the data, we believe the conclusions concerning main effects are valid.

Table 15. Ordinal Logistical Fit for Normalized Assessment Values

Level	Prob>ChiSq			
	Whole Model Test	Effect Likelihood Ratio Test		
		Scale	Simulation Belief	Scale cross Simulation Belief
Subtask	0.0001	0.0000	0.3706	0.0000
Task	0.0001	0.0000	0.9235	0.0000
Scenario	0.0001	0.0000	0.1792	0.0002
Overall	0.0001	0.0000	0.9604	0.0000

These results indicate the scale used can affect assessments and inter-SME consistency. The type of scale used by the rater also has the potential to mitigate the degree of inconsistency across SMEs and to produce inter-SME results that are both more consistent. Knowing there is inter-SME inconsistency, it is important to know whether the process is ‘in’ or ‘out of control’, if SMEs are biased, or some combination of process control and SME bias is affecting inter-SME consistency.⁶⁸

Based on the results described above concerning inter-SME consistency, it is apparent there is a need to develop methods to reduce variability in the process (where the process is the face validation of human behavior representation models). While this research primarily focuses on the effects of SME bias, this important finding will be addressed in the research agenda.

To help determine if the assessment process is in control, one must assess individual SME consistency.⁶⁹ Intra-SME *consistency* is a SME’s ability to maintain concurrence between the average of the sublevel response scores and the level score. To measure intra-SME consistency, the non-zero scores for each response are used to compare the differences between the mean sublevel value to the level value. Equation 1 is the general consistency score formula for sublevel-level pairings. Each sublevel-level pairing uses the general formula to calculate consistency scores. Each data point (C_{ij}) is the value of the difference between the given score for a level minus the mean value of

⁶⁸ An assessment process under control is one where the variance in responses is minimal resulting in consistent responses.

⁶⁹ A process out of control is a form of bias which introduces a systematic error into the sample by soliciting one response over another, disregarding information presented. *Process bias* occurs when a process’ procedures predispose results by affecting the ability of the SME to draw conclusions based on the evidence offered.

the non-zero ratings given to the sublevel associated with a given level for each individual SME rounded to the nearest whole number. To compare scores from different scales (7-Point, 5-Point, 2-Point (Go/No-Go)), the differences are normalized between -1 and 1. Each SME's data points, C_{ij} , are summed and divided by the total number of levels assessed by the SME to provide the SME's sublevel-level consistency score, C_i .

$$C_{ij} = \frac{1}{s} \text{round}(L_{ij} - \frac{1}{n} \sum_{k=1}^m E_{ijk})$$

Equation 1. Sublevel Consistency Score

$$C_i = \frac{1}{p} \sum_{j=1}^o C_{ij}$$

Equation 2. Level Consistency Score

where

C_{ij} : SME i 's normalized consistency score for sublevel question j

C_i : SME i 's normalized consistency score

E : Element (subtask, task, or scenario) score

L : Level (task, scenario, or overall) score

i : i^{th} SME

j : j^{th} level (task, scenario, or overall)

k : k^{th} element (subtask, task, or scenario)

m : Number of possible elements (subtasks, tasks, or scenarios) for level (task, scenario, or overall) j for SME i

n : $n \leq m$ and is the number of “non zero” response elements (subtasks, tasks, or scenarios) per level (task, scenario, or overall) j for SME i

o : Number of possible levels (tasks, scenarios, or overalls) for SME i

p : $p \leq o$ and is the number of “non zero” response for levels (tasks, scenarios, or overalls) for SME i

s : The normalization factor for the specified scale; 7-Point Likert Scale (7), 5-Point Likert Scale (5), and Go/No-Go Scale (2)

Table 16 shows the statistical likelihood of the factor being significant effect observing an effect based on the factors of scale and simulation belief at each sublevel-level pairing. The data is calculated using the absolute values of C_{ij} . Values of Prob>ChiSq less than 0.05 indicate a statistically significant effect of the factor.

The table denotes at least one factor is statistically significant for each sublevel-level pairing (Prob>ChiSq = 0.0001). Looking at the potential effect based on scale, the table indicates a statistically significant effect on consistency for all pairings (Prob>ChiSq = 0.0000). For simulation belief, no statistically significant effect is present. However, the scenario-overall pairing approaches significance with a Prob>ChiSq of 0.0589. Finally, there is no statistically significant effect or interaction based on scale cross simulation belief at any sublevel-level pairing.

Table 16. Ordinal Logistical Fit for Normalized Consistency Scores

Sublevel-Level Pairing	Prob>ChiSq			
	Whole Model Test	Effect Likelihood Ratio Test		
		Scale	Simulation Belief	Scale cross Simulation Belief
Subtask => Task	0.0001	0.0000	0.1857	0.6077
Task => Scenario	0.0001	0.0000	0.8727	0.6313
Scenario => Overall	0.0001	0.0000	0.0589	0.8575

Figure 32 shows the Sim-Scale Groups broken down into sublevel-level groups (x-axis) and the mean values of C_i (y-axis). Data indicates SMEs more harshly judge overall performance versus the scenario performance if told they are watching a constructive simulation rather than a playback of instrumented soldiers. No uniform pattern of increasing, decreasing, or steady assessment was displayed in the general tendencies of assessment based on group, scale, or simulation belief.

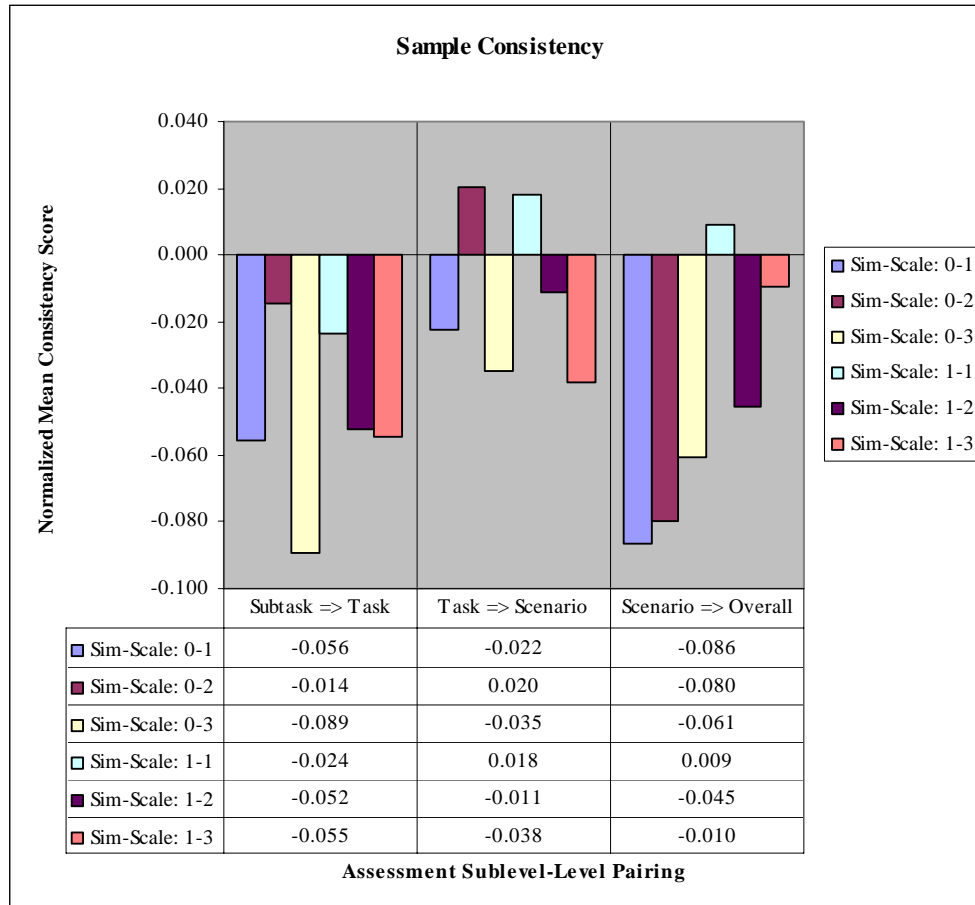


Figure 32. Intra-SME Mean Consistency Scores

Figure 33 graphically displays the lack of correspondence of the normalized, absolute value of the SMEs' mean subtask-to-task scores. The response (y-axis) is the absolute value of C_i for subtask and task ratings. The x-axis is the Sim-Scale Group. When grouped by scale, the mean consistency scores for the 5-Point Likert Scale (#-1) are greater than the mean consistency scores for the 7-Point Likert Scale (#-3). Mean consistency scores of the 7-Point and 5-Point Likert Scales are larger, less consistent, than the mean consistency scores of the Go/No-Go (#-2) Scale. The graphic illustrates the mean consistency scores based on simulation belief for the subtasks-task pairings do not indicate responses for SMEs who believe they are assessing human performance, live simulation (1-#), are any more consistent than scores from SMEs who believe they are assessing a constructive simulation (0-#). Figures N.3 and N.4 in

Appendix N. Supporting Figures and Tables for Data Analysis display the same trends for mean consistency scores when grouped by scale or by simulation belief for tasks-scenario and scenarios-overall pairings.

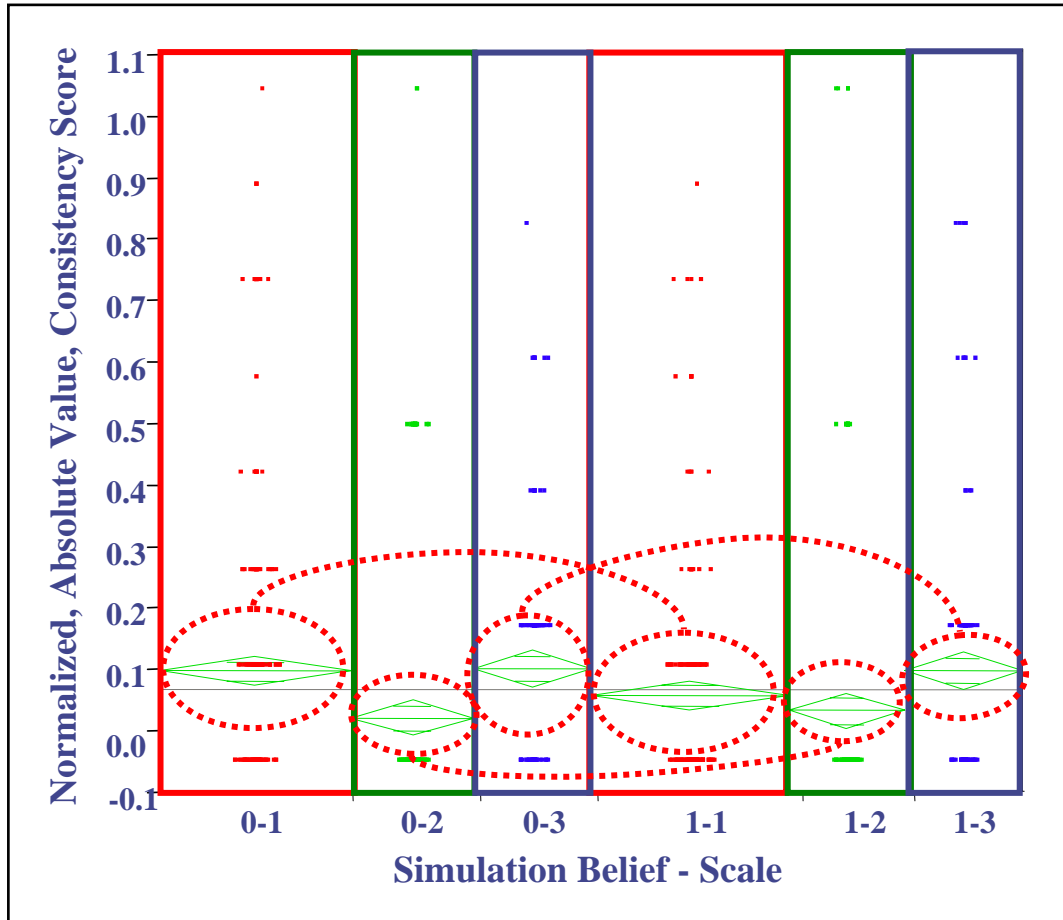


Figure 33. Intra-SME Subtask-to-Task Consistency Scores

Initial consistency analysis indicates mean SME assessments are inconsistent at all levels of interaction (subtask-to-task, task-to-scenario, scenario-to-overall, subtask-to-scenario, etc.) with an effect due to scale. However, does this affect the overall assessment of the model by changing the results between levels from Go to No-Go, No-Go to Go, Unknown to Go, etc? If this effect exists, it is an example of practical bias, the influence bias has on the assessment. The practical effect of inconsistency is referred to as *consistency impact*. Its measure is the percentage of sublevel-level pairing responses that change their assessment score based on consistency scores, valid versus invalid. The following is the general formula for calculating the impact of inconsistency:

$$M_{ij} = \frac{1}{n} \sum_{k=1}^m E_{ijk}$$

Equation 3. Mean Sublevel Consistency Score

$$B_{Mij} = \text{if } (M_{ij} > 0.667), \text{ then } 1, \text{ elseif } (M_{ij} > 0), \text{ then } (-1), \text{ else } 0$$

Equation 4. Binned Sublevel Consistency Score

$$B_{Lij} = \text{if } (L_{ij} > 0.667), \text{ then } 1, \text{ elseif } (L_{ij} > 0), \text{ then } (-1), \text{ else } 0$$

Equation 5. Binned Level Consistency Score

$$I_{ij} = \text{if } (B_{Mij} == B_{Lij}), \text{ then } 0, \text{ elseif } (B_{Mij} > B_{Lij}), \text{ then } 1, \text{ else } (-1)$$

Equation 6. Sublevel Consistency Impact Score

$$I_i = \frac{1}{p} \sum_{j=1}^o f(x_{ij})$$

Equation 7. Level Consistency Impact Score

where

I_{ij} : SME i 's consistency impact score for sublevel question j

I_i : SME i 's consistency impact score

B : Assessment score bin; “No-Go” = -1, “Not Applicable” or “No Opinion” = 0, and “Go” = 1

E : Elements (subtask, task, or scenario) score

L : Level (task, scenario, or overall) score

M : Mean element score

i : i^{th} SME

j : j^{th} level (task, scenario, or overall)

k : k^{th} element (subtask, task, or scenario)

m : Number of possible elements (subtasks, tasks, or scenarios) for task j for SME i

n : $n \leq m$ and is the number of “non zero” response elements (subtask, task, or scenario) per level (task, scenario, or overall) j for SME i

o : Number of possible levels (tasks, scenarios, or overalls) for SME i

p : $p \leq o$ and is the number of “non zero” response for levels (tasks, scenarios, or overalls) per SME i

Table 17 shows the possibility of effect on the consistency impact scores based on scale and simulation belief. The general consistency impact formula for I_{ij} produces the data used for this analysis. As with the calculations for consistency, values are calculated using the absolute value of I_{ij} in order to ensure the distribution of differences from all SMEs about zero do not excessively influence the potential degree of impact for inconsistency within SMEs. Values of Prob>ChiSq less than 0.05 indicate a lack of consistency impact for the level and interaction for the scale, simulation belief, or scale cross simulation belief. The table denotes an effect on consistency impact for each sublevel-level pairing, Prob>ChiSq is never greater than 0.0093. When looking at the potential effect based on scale, the table indicates a statistical effect on consistency impact for all sublevel-level pairings, Prob>ChiSq is always less than 0.0013. The table in Figure 34 indicates the Go/No-Go Scale, Scale 2, has the least effect on the consistency impact. For simulation belief, no effect is demonstrated, Prob>ChiSq is always greater than 0.4709. Finally, there is no statistical effect based on scale cross simulation belief (Prob>ChiSq is never less than 0.1896) for any sublevel-level pairing.

Table 17. Ordinal Logistical Fit for Normalized Consistency Impact Scores

Sublevels-Level Pairing	Prob>ChiSq			
	Whole Model Test	Effect Likelihood Ratio Test		
		Scale	Simulation Belief	Scale cross Simulation Belief
Subtask => Task	0.0001	0.0000	0.5422	0.2880
Task => Scenario	0.0093	0.0007	0.7154	0.8212
Scenario => Overall	0.0001	0.0000	0.6796	0.8912
Subtask => Task	0.0052	0.0013	0.9982	0.1896
Task => Scenario	0.0001	0.0000	0.7385	0.2702
Scenario => Overall	0.0001	0.0000	0.4709	0.9002

Figure 34 is a graphical display of the mean difference in consistency impact based on simulation belief and scale resulting from the consistency formula for I_i . The graphic indicates no general trends from sublevel-level pairing to sublevel-level pairing based on scale or simulation belief. Simulation-Scale Group 0-1 does show a consistent negative trend (-0.079 to -0.086) for mean impact consistency scores across sublevel-level pairings. This indicates approximately eight percent of level scores are harsher than their associated sublevel scores for the SMEs in Group 0-1.

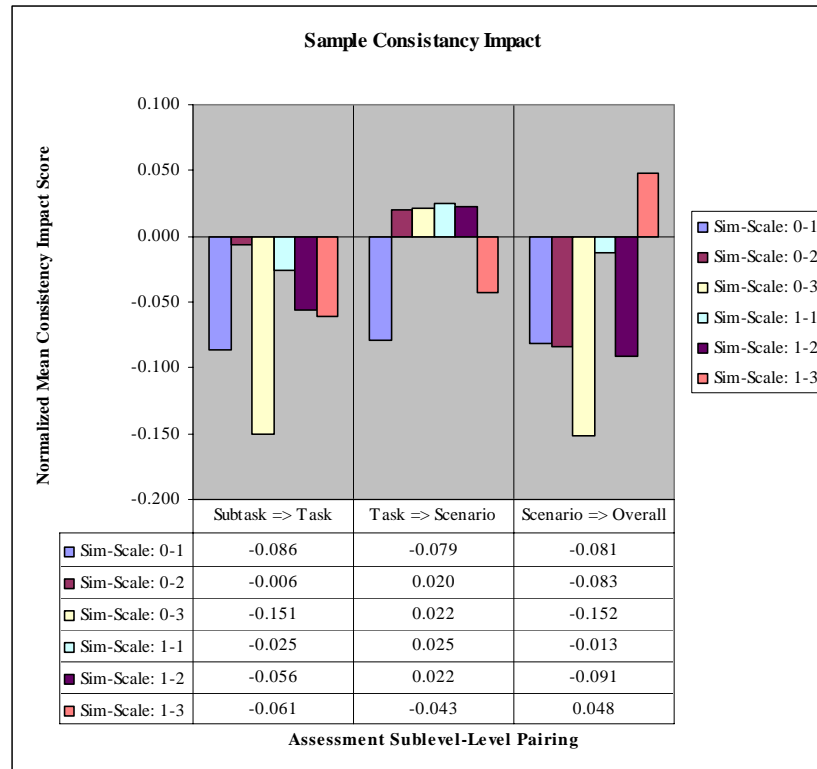


Figure 34. Intra-SME Mean Consistency Impact Scores

Although analyses of the mean values for the differences between the sublevel-level pairing assessments show no consistent pattern in general (process) or practical, a question remains regarding process accuracy. For this research, *accuracy* is defined as the rater's ability to maintain relative correctness with respect to a consistent, scale-dependent, assessment key for each subtask, task, scenario, and overall assessment.

Accuracy is measured using the normalized (-1 to 1) differences between the base assessment and SME assessments. These values are calculated using the following general accuracy formula

$$A_{ij} = \frac{1}{s}(E_{ij} - B_{ij})$$

Equation 8. Sublevel Accuracy Score

$$A_i = \frac{1}{n} \sum_{j=1}^m f(x_{ij})$$

Equation 9. Level Accuracy Score

where

A_{ij} : SME i 's normalized accuracy score for question j

A_i : SME i 's normalized accuracy score

B : Baseline (key) score (subtask, task, scenario, or overall)

E : Element (subtask, task, scenario, or overall) score

i : i^{th} SME

j : j^{th} level (subtask, task, scenario, or overall) response

m : Number of possible level (subtask, task, scenario, or overall) responses for SME j

n : $n \leq m$ and is the number of "non zero" (subtask, task, scenario, or overall) responses j for the SME

s : The normalization factor for the specified scale; 7-Point Likert Scale (7), 5-Point Likert Scale (5), and Go/No-Go Scale (2)

Assessment questions that ask SMEs to forecast model performance in different environments, scenarios, or situations are not included in the accuracy assessment. Only

the last two assessment questions dealt with forecasting the capability of the HBR to perform doctrinally correct in unobserved situations. These two questions are predictive and qualitative in nature.

Table 18 shows the statistical likelihood of effect on accuracy based on the terms of scale and simulation belief for each level of assessment. As with consistency and consistency impact, the data is calculated using the absolute values of A_{ij} . A statistically significant effect is found at each level of assessment ($\text{Prob}>\text{ChiSq} < 0.05$). Looking at effect based on scale, the data indicates a statistically significant effect on accuracy for all levels, $\text{Prob}>\text{ChiSq}$ is always less than 0.05. For simulation belief, no statistically significant effect is present except at the overall assessment level with a $\text{Prob}>\text{ChiSq}$ of 0.0017. Finally, except for the subtask assessment level, $\text{Prob}>\text{ChiSq}$ of 0.0007, there is no statistically significant effect based on scale cross simulation belief.

Table 18. Ordinal Logistical Fit for Normalized, Absolute Value, Accuracy Scores

Level	Prob>ChiSq			
	Whole Model Test	Effect Likelihood Ratio Test		
		Scale	Simulation Belief	Scale cross Simulation Belief
Subtask	0.0001	0.0000	0.1151	0.0007
Task	0.0378	0.0177	0.4687	0.4016
Scenario	0.0003	0.0006	0.5831	0.0572
Overall	0.0001	0.0000	0.0017	0.6216

Figure 35 presents the intra-SME assessment accuracy using the mean values of A_i . The data indicates that although all the scales start out relatively tight in their mean accuracy scores at the subtask level, they diverge in their degree of accuracy at successive levels of assessment.⁷⁰ SMEs using the Go/No-Go Scale rated performance more harshly at the subtask level and more leniently at subsequent levels than the key assessment or SMEs using the other two scales.

⁷⁰ Note: subtasks show specifics, tasks are general, scenarios are specific situations for a specific environment, and the first two overall assessment questions are for a specific environment.

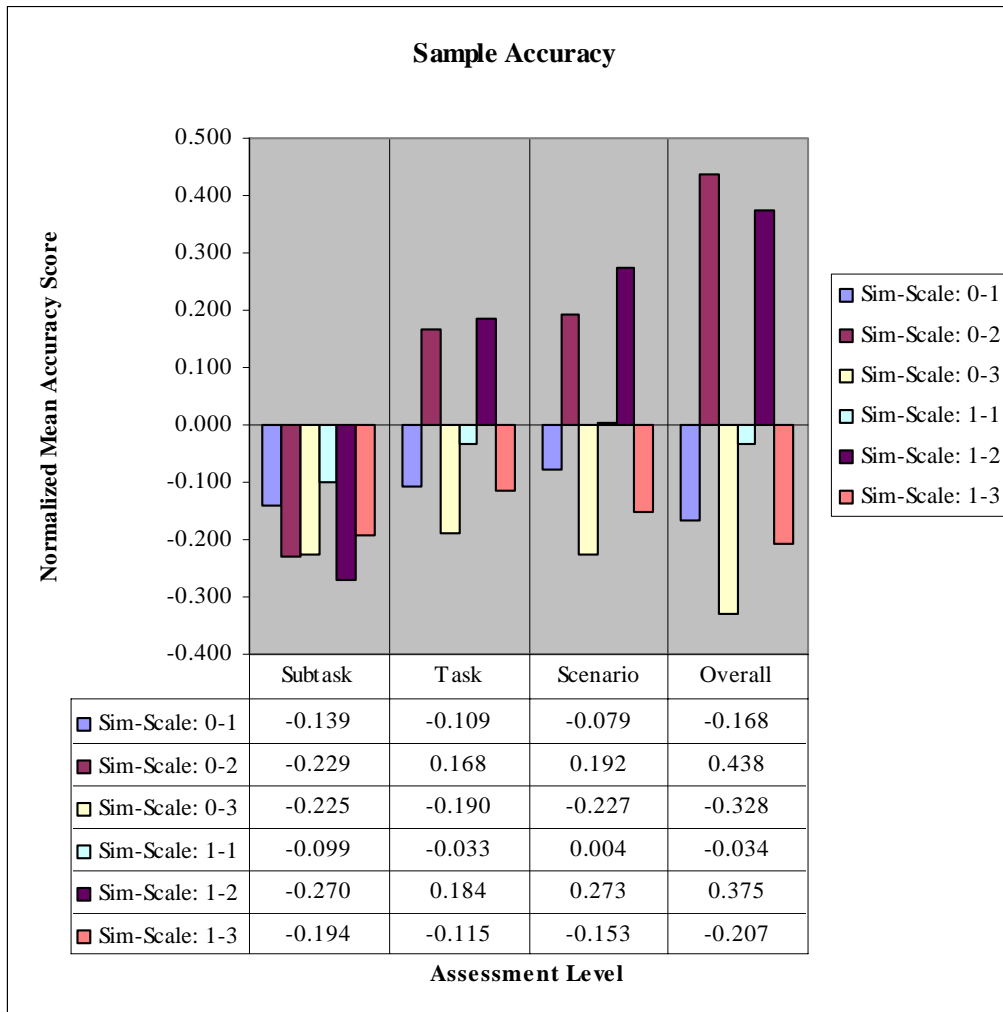


Figure 35. Intra-SME Mean Accuracy Scores

Analyses of the accuracy of SME responses shows an effect based on scale, but this begs the question of the practical impact of this inaccuracy. *Accuracy impact* is the affect inaccuracy has on the general assessment of the subtask, task, scenario, or overall performance. It shows the percentage of questions that differ in their relative value based on differences in accuracy scores, “Go” versus “No-Go”. Similar to the relationship between consistency and consistency impact, accuracy impact is the percentage of level responses that differ in overall assessment score based on the response’s accuracy score, valid versus invalid. Accuracy impact is calculated using the following general accuracy formula:

$$N_{ij} = \frac{1}{s} E_{ij}$$

Equation 10. Normalized Element Score

$$B_{Nij} = \text{if } (N_{ij} > 0.667), \text{ then } 1, \text{ elseif } (N_{ij} > 0), \text{ then } -1, \text{ else } 0$$

Equation 11. Binned Element Score

$$B_{Kij} = \text{if } (K_{ij} > 0.667), \text{ then } 1, \text{ elseif } (K_{ij} > 0), \text{ then } -1, \text{ else } 0$$

Equation 12. Binned Assessment Key Score

$$J_{ij} = \text{if } (B_{Nij} == B_{Kij}), \text{ then } 0, \text{ elseif } (B_{Nij} > B_{Kij}), \text{ then } 1, \text{ else } -1$$

Equation 13. Sublevel Accuracy Impact Score

$$J_i = \frac{1}{n} \sum_{j=1}^m J_{ij}$$

Equation 14. Level Accuracy Impact Score

where

J_{ij} : SME i 's normalized accuracy impact score for question j

J_i : SME i 's normalized accuracy impact score

B : Assessment score bin; “No-Go” = -1, “Not Applicable” or “No Opinion” = 0, and “Go” = 1

E : Element (subtask, task, scenario, or overall) raw score

K : Key (subtask, task, scenario, or overall) raw score

N : Element (subtask, task, scenario, or overall) normalized score

i : i^{th} SME

j : j^{th} level (subtask, task, scenario, or overall) response

m : Number of possible level (subtask, task, scenario, or overall) responses for SME i

n : $n \leq m$ and is the number of “non zero” (subtask, task, scenario, or overall) responses j for SME i

s : The normalization factor for the specified scale; 7-Point Likert Scale (7), 5-Point Likert Scale (5), and Go/No-Go Scale (2)

Table 19 shows the statistical likelihood of effect on accuracy impact based on the terms of scale and simulation belief for each level of assessment. As with consistency, consistency impact and accuracy, the data is calculated using the absolute values of J_{ij} . The table denotes an effect at each level of assessment (Prob>ChiSq = 0.0001). Looking at the potential effect based on scale, the table indicates a statistical effect on accuracy impact based on scale for all levels (Prob>ChiSq = 0.0000). For simulation belief, a statistically significant effect is present at the subtask and task level with a Prob>ChiSq of 0.0006 and 0.0024 respectively. Finally, except at the overall assessment level, Prob>ChiSq of 0.1216, there is a statistically significant effect based on scale cross simulation belief.

Table 19. Ordinal Logistical Fit for Normalized Accuracy Impact Scores

Level	Prob>ChiSq			
	Whole Model Test	Effect Likelihood Ratio Test		
		Scale	Simulation Belief	Scale cross Simulation Belief
Subtask	0.0001	0.0000	0.0006	0.0101
Task	0.0001	0.0000	0.0024	0.0029
Scenario	0.0001	0.0000	0.0629	0.0381
Overall	0.0001	0.0000	0.3074	0.1216

Figure 36 is created using the values of J_i for accuracy impact. The graphic indicates there are no general trends from assessment level to assessment level based on scale or simulation belief. Group 0-3 illustrates a trend toward increasingly less accurate responses at each level of assessment. Although the accuracy showed a trend for SMEs using the Go/No-Go Scale to become more lenient in their assessment with each successive level, the impact of the increasing leniency is to keep the assessment slightly negative (between -0.0333 and -0.2000) for the task, scenario, and overall assessment levels. When SMEs used the 5-Point Likert Scale, the scores get progressively harsher

from task to scenario to overall assessment level even though the accuracy graph shows accuracy for the other two scales maintaining a relatively constant negative value across all four levels of assessment.

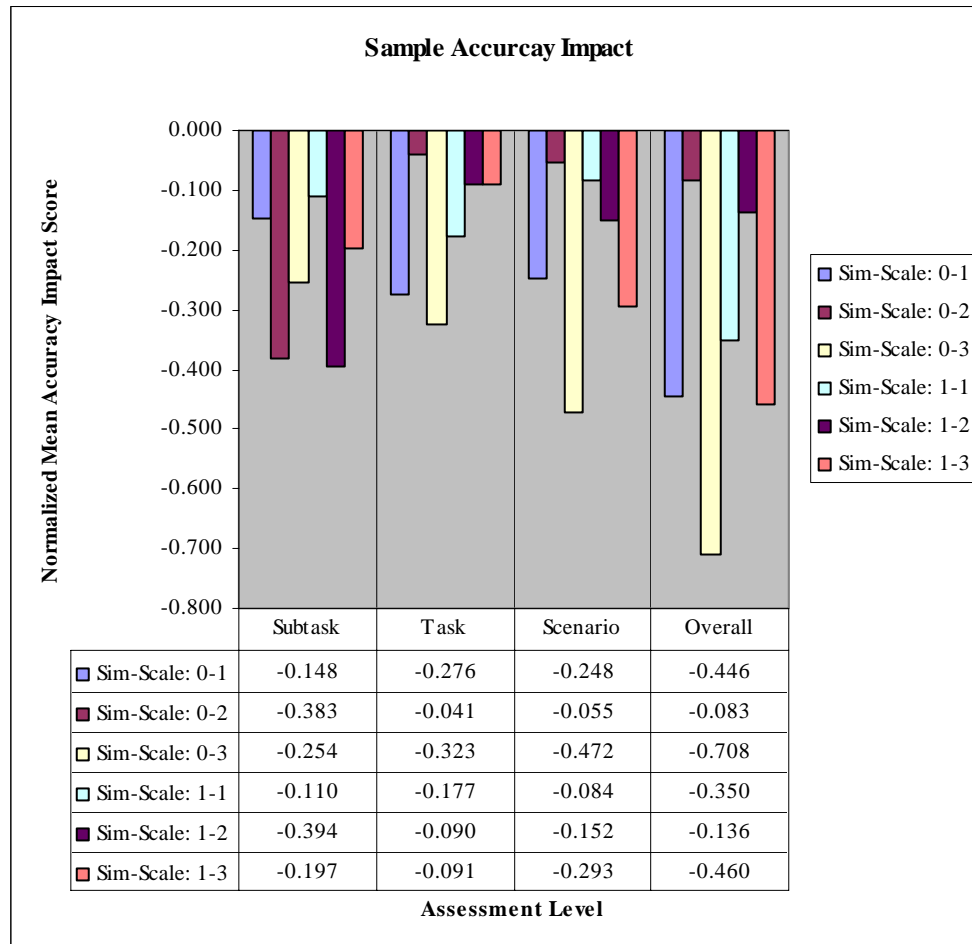


Figure 36. Intra-SME Mean Accuracy Impact Scores

The effect of bias on consistency and accuracy is shown through the removal of responses from SMEs who demonstrated patterns of bias. Table 20 shows the result of the overall assessment scores by group after 97 SMEs (53.30%) demonstrating one or more of the four identified bias are removed. All but one of the twenty-eight cells increased their mean value score. Due to this general increase in the assessment scores, six of the mean scores changed from “No-Go” to “Go”. This indicates a decrease in consistency for the mean cell response but results in a higher inter-SME general assessment consistency.

Table 20. Normalized, Mean Overall Assessment Scores - Minus Bias

ID		Number of SMEs	Mean (Normalized 0-1 Responses)			
Simulation Belief	Scale		Overall 1	Overall 2	Overall 3	Overall 4
0	1	16	<i>0.589</i>	<i>0.598</i>	<i>0.563</i>	<i>0.580</i>
0	2	21	1.000	1.000	1.000	1.000
0	3	7	<i>0.543</i>	<i>0.543</i>	<i>0.514</i>	<i>0.543</i>
1	1	16	0.777	0.768	0.696	0.714
1	2	15	0.967	1.000	0.900	0.933
1	3	10	0.700	0.700	<i>0.660</i>	<i>0.660</i>
All Beliefs and Scales		85	0.802	0.808	0.763	0.778

Figure 37 illustrates the effect on inter-SME consistency for overall model assessment questions when those SMEs identified as having bias are removed from the sample. As with Figure 29, the x-axis is the normalized raw response score and the y-axis displays the number of SMEs who assessed the question, Overall 1, at the specific score. Scores ranged between 0 and 1 a the mean of 0.802. The median (0.857) falls to the left of the means diamond and the responses are not normally distributed. Thus, by removing SMEs who were identified as biased, the standard deviation is reduced to 0.248, outliers are identifiable, the percentage of SMEs judging the question as a “Go” increased to 81.18%, and inter-SME consistency is enhanced.

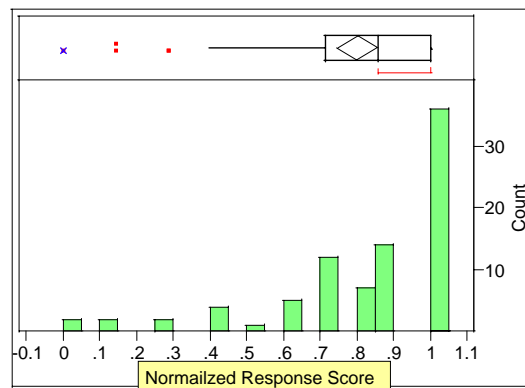


Figure 37. Distribution of Subject Matter Experts' Normalized Responses to Question Overall 1 without Biased Subject Matter Expert Responses

The general effect on intra-SME accuracy impact when excluding SMEs demonstrating bias indicates, except for Group 1-3, accuracy impact increases for the

task, scenario, and overall assessment levels (Appendix N., Table N.17).⁷¹ At the subtask level, those using the 7-Point Likert Scale accuracy impact increased. For groups using the 5-Point Likert or Go/No-Go Scales, the accuracy impact decreased at the subtask level. Accuracy increased by as little as 0.87% and as much as 100% for 18 of the 24 level and group cells, while decreasing by 2.41% to 87.50% for the remaining six cells. The composite mean accuracy score increased from -0.3721 to -0.1882 improving the accuracy score by 49.41%.⁷²

In summary, there is a difference in the magnitude of inter-SME consistency and intra-SME consistency, consistency impact, accuracy, and accuracy impact based on the scale used. In general, SMEs using the Go/No-Go Scale are more consistent, have less inconsistency impact at the upper two sublevel-level pairings, and less inaccuracy impact at the task, scenario, and overall levels of assessment. However, SMEs who used the 7-Point Likert Scale have less inconsistency impact at the subtask-task pairing, are more accurate across all levels of assessment, and have less inaccuracy impact at the subtask level. SMEs using the 5-Point Likert are the least consistent, have the highest consistency impact scores, are less accurate, and have the greatest accuracy impact scores. Finally, the removal of responses from SMEs who demonstrate patterns of performance, anchoring, contrast, or confirmation bias increases the overall consistency and accuracy of the remaining SME response scores.

D. STEPWISE LOGISTICS REGRESSION

Using SME demographic factors and accuracy impact scores a forward stepwise logistic regression was performed to identify terms which demonstrated the greatest potential of having a statistically significant effect on accuracy impact scores. An ordinal logistical regression model was created using the 20 terms identified. Table 21 is the result of the ordinal logistic fit for the new mode. The data indicates 12 terms, 10 unique, have statistically significant effect on accuracy impact.

⁷¹ As mean scores approach zero, accuracy impact “increasing”. As mean score diverge from zero, accuracy impact “decreases”.

⁷² This score is calculated using each SME’s mean accuracy impact score.

Table 21. Term Estimates from Ordinal Logistic Fit for Accuracy Impact Scores⁷³

Terms	Estimate	Chi-Square	Prob>ChiSq
Intercept [1]	-0.0464	0.6600	0.4170
Scale {3&1 - 2}	0.3645	523.9800	<.0001
Scale {3 - 1}	-0.1477	64.4700	<.0001
Infantry [-1]	0.1441	34.7800	<.0001
Interests Quad: Introspectors [-1]	-0.1704	29.1600	<.0001
Performance Bias [1-0]	-0.2488	24.5300	<.0001
Time Since in Last Unit (> 6 Months) [-1]	-0.0658	16.9500	<.0001
First Shooter Video Experience {Expert - Average & None & Novice}	0.1115	14.4700	0.0001
First Shooter Video Experience {Average & None - Novice}	0.0517	10.0500	0.0015
Duty Position: Executive Officer [-1]	0.0422	8.4200	0.0037
Interactions Quad: {A&AB&AC&C - B&BD&CD&D} [-1]	0.0561	6.9000	0.0086
Duty Position: Squad Leader [-1]	0.0553	6.4500	0.0111
NEO-FFI: Neuroticism (3&4&5 - 1&2) [-1]	-0.0397	5.3600	0.0207
First Shooter Video Experience {Average - None}	0.0303	2.8200	0.0934
Duty Position: Scout Platoon Leader [-1]	-0.0318	2.7500	0.0972
Activity Quad {A&AB&AC&B&C - D&BD&CD} [-1]	-0.0373	1.8400	0.1752
Impulse Control Quad {A&AB&AC - B&BC&C&CD&D} [-1]	0.0347	1.2900	0.2563
NEO-FFI: Conscientiousness (3&4&5 - 1&2) [-1]	-0.0221	1.2000	0.2733
Total Time Service (10 Years) [0]	0.0195	1.0200	0.3123
Anger Control Quad {A&D - AB&AC&B&BD&C&CD} [-1]	-0.0140	0.4500	0.5034
Defense Quad: Maladaptive [-1]	-0.0075	0.0600	0.8144

The term with the largest effect on the R^2 value for accuracy impact scores is *Scale* with a Chi-Square of 523.98, which is, more than 15 times greater than the next largest unique term. Averaging each SME's accuracy impact scores, the Go/No-Go Scale (Scale 2) is the least accurate; however, SMEs who use this scale were more accurate and less harsh at the task, scenario, and overall levels than SMEs using the other two scales (Figure 37). The 7-Point Likert Scale (Scale 1) is the most accurate of the three assessment scales.

A SME's *Branch*, whether or not a SME is an Infantry officer, had the second largest effect on accuracy impact. As expected, scores indicate Infantry officers are more consistent as a group; however, internally they are less accurate in their appraisal of the observed behaviors. This indicates that although specific assessment criteria for subtasks and tasks are provided, experience gained from training, executing, and assessing these

⁷³ The absolute value of the accuracy impact score was used for this analysis.

tasks results SMEs with stronger opinions on what is doctrinally correct behavior. These individuals used the outermost values of the scale and consequently, when they differed from the key assessment score, the difference was greater than if they were less confident in their responses with responses hovering around the midpoint of the scale. Non-Infantry officers tended to judge behaviors in accordance with the prescribed face validation methodology and held less conviction of what they felt was doctrinally correct behavior, i.e. their responses were closer to the midpoint of the scale.

Although the stepwise comparison identified eight personality traits with a potential effect on accuracy impact scores, only three show statistical significance: neuroticism scores, interaction style, and learning style.⁷⁴ The *Interests* style quadrant categories ‘D&BD&CD’ are a combination of a SME’s general lack of extraversion and a positive openness score.⁷⁵ The NEO-FFI classifies these SMEs as “Introspectors”.⁷⁶ SMEs with reported “Introspectors” personality styles had poorer accuracy impact scores than those who reported a tendency towards being more social or whose interests are more conventional. The *Interaction* style categories ‘A&AB&AC&C’ are a combination of SME extraversion and lower agreeableness scores. The NEO-FFI classifies these SMEs as “Leaders” or “Competitors”.⁷⁷ SMEs with these reported personality styles tended to have less accurate and harsher scores than SMEs who reported a tendency towards being more concerned with others than with themselves. Of the three personality terms demonstrating a statistically significant effect on accuracy impact, neuroticism scores show the least effect, Chi-Square = 5.36. SMEs with lower neurotic scores have slightly decreased accuracy impact scores and a tighter distribution of results; thus, they are more accurate than SMEs with higher neurotic scores. These results indicate SMEs who

⁷⁴ As discussed previously, the NEO-FFI identifies and describes ten personality styles based on the pair-wise interaction of the five NEO-FFI score categories.

⁷⁵ Quadrants are labeled as: A - upper left; B – upper right; C – lower left; and D – lower right. Labels such as AB, AC, BD, and CD indicate scores on the boundary between two quadrants, i.e. AB is a score on the boarder of quadrant A and quadrant B.

⁷⁶ “Introspectors” seek activities that they can perform alone. However, “Introspectors” generally have higher levels of imagination and prefer challenging activities. [COST 00]

⁷⁷ In general, “Leaders” are social individuals who are adept at making decisions and prefer issuing instructions. “Competitors” are individuals who possess the same general agreeableness traits as leaders but are distrustful of others and favor keeping to themselves. [COST 00]

display a propensity towards being more conventional, more agreeable, and less neurotic provide more consistent, less harsh, and more accurate assessment of HBR models that replicate dismounted infantry tasks in an urban environment.

Performance bias is the term showing the fourth greatest effect on accuracy impact for the sample population. The effect indicates those demonstrating performance bias are internally less accurate and less consistent than those who do not exhibit performance bias. This is expected since performance bias is identified in the data as SMEs who fail to provide responses. Since any SME's response has a better chance of being accurate and consistent than no response, excluding personnel who demonstrate performance bias will result in more accurate and more consistent validation results.

Time Since in Last Unit is broken down into two groups: SMEs with more than six months and SMEs with six months or less since last with troops. Table 21 shows a statistically significant effect based on this term. Participants away from troops for more than six months are more accurate in their assessment of observed behaviors than those who had been away from troops less than six months. This is likely due to the SMEs being less opinionated and more tolerant of alternative methods since they have not recently been performing or assessing soldiers performing these tasks.

Participants with more experience playing first shooter video games are more accurate and harsher in their assessment of observed behaviors than SMEs with less experience playing first shooter video games. This may be due to their ability to extract more information in a shorter period of time from a computer display or their increased familiarity with simulation and general acceptance of computer technology.

Participants who held the duty positions of squad leader (SL) or company executive officer (XO) were less accurate and harsher in their assessment of observed behaviors than those who held neither of these positions. A squad leader is an enlisted soldier's duty position and indicates extended prior enlisted time. It normally takes five to seven years of service before an enlisted soldier has the training, experience, and opportunity to hold the position of squad leader. Commanders typically assign junior officers who have proven themselves as platoon leaders to the position of executive

officer, which suggest these personnel served at least a year or two in line units in leadership positions. Both duty positions, SL and XO, normally require personnel to demonstrate proficiency at other duty positions before commanders assign them one of these duty positions. As with Infantry soldiers, SMEs who had held SL or XO positions were more opinionated and prone to using the outermost scale values than SMEs who had not held these duty positions.

The stepwise logistical regression did not identify simulation belief as one of the top 20 terms with potential for a statistically significant effect on accuracy impact scores. This reinforces the hypothesis that there is no statistically significant effect on SME responses based on simulation belief.

THIS PAGE INTENTIONALLY LEFT BLANK

V. DISCUSSION

From analysis of results of the two base studies, numerous insights are exposed. This chapter discusses five of these: the effects of simulation belief and use of MTP assessment forms, the role of assessment scales, effects of bias, effects on validation criteria, and the impact of personality on bias.

A. SIMULATION BELIEF

The null hypothesis for the first study stated that assessment of human performance shows no difference with regards to bias for SMEs who believe they are assessing simulated behaviors and those who believe they are assessing real-world behaviors using conventional validation methods. Analysis of the data indicates SMEs using the 7-Point Likert Scale demonstrated the same number of bias cases whether they believed they were assessing simulated behaviors or human behaviors. This means we fail to reject the null hypotheses and conclude that we can use the same MTP evaluation checklist to assess human performance and HBR performance of the same ground combat urban operation tasks. Furthermore, there was no statistically significant effect of simulation belief combined with the consistency, consistency impact, accuracy, and accuracy impact scores giving further weight to this conclusion.

The tasks, scenarios, and overall assessment questions represent MOEs for the validation process. Subtask questions represent the MOPs for the validation process. A fundamental assumption is that MTP worksheets for assessing the performance of a soldier executing these tasks provide valid criterion for individual soldiers and teams to perform a task. Provided this assumption is valid, similar checklists should be developed for other types and levels of human behavior. These new checklists would need to be verified for applicability as referent and viability as appropriate assessment standards.

B. ASSESSMENT SCALES

The null hypothesis for the second study stated SMEs demonstrate the same levels of effect on consistency and accuracy during validation of an HBR model implementation using a 7-Point Likert Scale as they do when using a 5-Point Likert Scale or Go/No-Go Scale. Analysis of the data indicates SMEs using the Go/No-Go Scale were more

consistent and accurate at the task, scenario, and overall levels of assessment. However, SMEs using the 7-Point Likert Scale were more accurate and consistent at the subtask to task level of assessment. This means we reject the null hypothesis and accept the alternative hypothesis that scale has an effect on the magnitude of intra-SME consistency, consistency impact, accuracy, and accuracy impact.

The effects of scale indicate conflicting results. In one case, the fewer grading options presented to a SME, Go/No-Go, the tighter the consistency and accuracy impact scores. However, in the comparison of the 7-Point Likert and 5-Point Likert Scales, the more options presented the SME for possible responses, the tighter the consistency and accuracy impact scores.

The Go/No-Go Scale forces SMEs to place observed actions into two categories with no middle ground for questionable performance. The 7-Point Likert Scale enables SMEs to make finer discrimination of observed behavior on the scale from very poor to excellent. The two-point scale provides more accuracy and consistency because it forces SMEs to make a black and white decision. Although all three scales afford SMEs the opportunity to comment on their assessments, the two-point scale fails to provide SMEs with the ability to quantitatively express the extent to which they felt the HBR model's performance was or was not to standard. Thus, the 7-Point and 5-Point Scales provide a quantitative method for SMEs to make fine-grained assessments to more accurately rate the observed behaviors. Consequently, SMEs can paint a truer picture of their assessment of the capabilities and limitations of the model using the Likert Scales versus the Go/No-Go Scale.

Quantitative measures provide a more effective means of expressing the validity of the model. Results based on quantitative measures are difficult to dispute without further investigation or assessment of the model. Qualitative assessments, such as SME comments, are difficult to correlate and summarize. However, qualitative responses can help clarify issues identified in the quantitative data. For the best results, a mixture of quantitative scales to provide data for more traditional statistical analysis and space for qualitative responses to amplify SME assessments provides the best chance of providing a comprehensive validation. A Go/No-Go Scale should be used at the lowest level of

assessment or when the number of sublevel components exceeds seven to reduce the mental requirements for tracking scores and increase consistence between the subtask and task level of assessment.⁷⁸ At the aggregate levels of assessment, a 7-Point Likert Scale would be best in order to provide SMEs the ability to express the degree of proficiency a task was performed or a scenario was executed.

C. BIAS

Bias has been shown to affect the accuracy and consistency of SME responses making face validation results less conclusive and reliable. The next question is how to mitigate their effect on face validation results.

Confirmation bias showed an effect on intra-SME consistency and consistency impact. Inconsistency manifested in SME assessments through confirmation bias may be resolved using SME comments. This requires a SME to recognize one or more subtask(s)/task(s)/scenario(s) influence his final assessment more than others and comment to this fact. This identification and comment practice rarely manifested itself in this research.

One potential system for identifying and mitigating confirmation bias during assessment is a computerized tool which supports SMEs and validation agents in face validation efforts. A computer assisted assessment system is discussed in the next section, Validation Criteria. A second technique is to weight each assessment question to provide SMEs the ability to identify factors that are of greatest importance in the assessment of the HBR model's behaviors.

Anchoring and contrast bias have proven to have an effect on the accuracy of the SMEs responses. Inaccuracies, which manifested themselves in SME assessments through these biases, can be resolved using SME comments. However, this information is qualitative in nature and is difficult to quantify to alleviate inaccuracies. As was the case with confirmation bias, SME comments were routinely lacking. One method to mitigate

⁷⁸ Seven components is recommended due to the ability of the average individual to track five to nine factors at anyone time [MILL 56].

anchoring and contrast bias is to affix each question's scale options with well-defined descriptions and examples to provide SMEs with more exacting criteria for inadequate, ambiguous, and appropriate performance.

Performance bias affects both accuracy and consistency. The validation agent can mitigate a willing SME's inability to comply with validation procedures through additional training and the use of specific textural and visual examples of poor, fair, and excellent task performance. These well-defined examples could help alleviate SME questions on the standards for proper performance. Coupled with retraining and additional practice sessions for SMEs identified as having difficulty with the assessment process, providing specific examples could alleviate some aspects of performance bias.

The validation agent may identify SMEs who possess or develop an uncooperative attitude toward the validation process during the evaluation of training and practice sessions. This unaccommodating attitude can be addressed either through counseling of the SME or the removal of the SME from the process. SMEs that possess an uncooperative attitude may still go through the validation process and their responses could be used to examine potential qualitative insights.

Additional training can allow the SME pool to obtain and maintain a level of proficiency in the validation process that it failed to possess prior to the training. Training and practice sessions help to identify SMEs with the potential for each type of bias and provided an opportunity to mitigate bias through further training or process modifications.

D. VALIDATION CRITERIA

In general, validating agents can enhance SME consistency and accuracy through the training of SMEs. Participants for this research went through one hour of training on the validation procedures. As described in Sections IV.B. Bias Patterns, a more extensive training program with clear visual and textural descriptions of poor, questionable, and exceptional performance would help to reduce the lack of consistency and accuracy. Acknowledging the time allowed for training SMEs on the validation process will always

be limited, the remainder of this section addresses how the validation agent can modify the assessment worksheets and the face validation procedures to address issues of consistency and accuracy.

Inter- and intra-SME inconsistency exists at sublevel-level pairings of assessment. These inconsistencies make it difficult, if not impossible, to provide an accurate assessment of the model. One method of decreasing the amount of inter-SME inconsistency is to mitigate intra-SME inconsistencies. Mitigating intra-SME inconsistency will enhance the face validation process by helping produce conclusive results.

Although numerous factors may contribute to a SME demonstrating inconsistency between sublevel assessment responses and level assessment responses, the validation process can institute two possible techniques to help resolve the issue: computerized assessment process and specific, fixed scale criteria for each assessment question.

One factor adding to SME inconsistency is the number of sublevel assessments. Numerous sublevel questions per level makes it difficult for SMEs to mentally tally and track the mean sublevel score. Providing a computerized system to calculate intra-SME consistency and warn the SME of potential inconsistencies could alleviate the need for SMEs to track their sublevel scores. Once identified, the system should allow the SMEs to provide justification for inconsistencies or modify their responses to mitigate inconsistencies. The same system, while tracking intra-SME consistencies, can calculate inter-SME inconsistency. The computer software would provide an inter-SME consistency report to the validation agent who can investigate and deconflict any issues.

Another method to mitigate SME inconstancy is to allow SMEs to place a weighting factor on each sublevel response they feel affects the level assessment to a greater or lesser degree. This allows SMEs to identify what they consider the more important issues. Assessment question weighting factors increase consistency by allowing the mean of the sublevel assessments to correlate more closely with the assessment value of the level. This helps to ensure the whole is a reflection of the parts.

There also exists an issue of intra-SME accuracy and accuracy impact at each level of assessment. Accuracy was measured based on the difference of the SMEs responses from an assessment key scale. This allowed the comparison of SME scores relative to their difference to a consistent baseline assessment. One method of increasing accuracy is to provide SMEs with more exacting descriptions for Likert Scale responses. Grounding assessment scales by providing specific descriptions for each response option is a method used by human resource personnel to enhance the assessment process of employees [CHAR 02] [DRUK 88] [GAWR 00] [STUF 02]. The Likert Scales for this research used generalized values and a comments section for each question to provide SMEs with the freedom to express their view of model performance. This is meant to limit possible effect of bias introduced by a validating agent who might over focus SMEs by providing more specific scale score definitions.⁷⁹

There are two means for grounding assessment scales. The first method fixes values for the tails of the scale, *general grounding*. The second method is to ground each scale value for each question, *explicit grounding*.

General grounding allows the validation agent to fix the boundaries of the assessment scale while affording SMEs flexibility to judge questionable actions based on their experiences. To provide detailed descriptions for general grounding requires extensive knowledge acquisition and referent validation. Although the process fixes the extremes, it still allows imprecise responses about the scale's median score.

Explicit grounding allows the validation agent to fix the internal scale values as well as the boundaries of the assessment scale for each question. The process can make judgment of borderline and boundary behaviors more accurate between SMEs. Care must be taken not to limit the ability of SMEs to use their experience to provide insight into the assessment of the HBR model. SMEs may still provide comments to address issues they have with the model's performance, but these may be lost in analysis of the quantitative responses to the assessment process. Explicit grounding requires more

⁷⁹ Over focusing is a form of framing bias. Framing bias occurs when SMEs are unaware of the focus of their efforts and thus they are assessing things the model is not intended to address. However, framing bias can also occur if a SME is so focused that he is not accounting for peripheral factors that influence model performance or its potential functionality.

extensive knowledge acquisition and referent validation then general grounding, thus, requiring the expenditure of more man-hours and dollars to produce. This may not be a viable option if funding is limited or time constraints are prohibitive.

Grounding Likert Scale scores for each question provides a means of increasing accuracy by providing SMEs with specific examples of behavior scale scores. Validating agents may also wish to provide visual examples of the different behaviors to ensure SMEs have similar mental image of the behavior criteria.

E. BIAS AND PERSONALITY

To the extent it influences bias, validating agents can use personality styles to identify and potentially mitigate the effects of bias on SME consistency and accuracy. The link between bias and SME personality holds the possibility for identifying one of the underlying motivations for the differences in SME observations. The remainder of this section illustrates the influence of personality, as categorized through NEO-FFI results, on the four distinct bias types examined in this research. The findings come from stepwise logistics regression and ANOCAT on the five NEO-FFI score values and the NEO-FFI personality styles.

Analyzing the interaction between confirmation bias and neuroticism, *Interaction*, and *Interest*, demonstrates a statistically significant effect based on *Interest*. “Introspectors” demonstrate a higher propensity of displaying confirmation bias than the other three personality categories associated with the *Interest* personality style.⁸⁰ Concerning *Interaction* style categories, SMEs who display a tendency towards being more concerned with others than with themselves are less likely to display confirmation bias.

Further analysis of effects based on personality styles, identifies an additional style demonstrating statistically significant effect on bias, *Attitudes*. Participants who

⁸⁰ The four categories within the *Interest* personality style are “Mainstream Consumers” (greater extraversion and less openness), “Creative Interactors” (greater extraversion and greater openness), “Home Bodies” (less extraversion and less openness), and “Introspectors” (less extraversion and greater openness) [COST 00].

display confirmation bias also show less of the attitude characteristic of “Result Believers”.⁸¹ Results indicate individuals who demonstrate innovative non-traditional thought also show less confirmation bias.

Analyzing the interaction between contrast bias and personality terms, the terms neuroticism, *Interaction*, and *Interest*, demonstrate no statistically significant effect. Further analysis shows no other personality terms show a statistical significance effect with respect to contrast bias. This is likely do to only five SMEs demonstrating contrast bias. Based on these results, there are no inferences made with regard to personality and contrast bias.

Performance bias, neuroticism, the personality style *Interaction*, and the personality style *Interest* have an effect on accuracy impact. Analyzing the interaction between performance bias and these personality terms reveals no statistically significant effect. Additional analysis on the other personality styles indicates there is a statistically significant effect on performance bias based on *Learning*, *Activity*, and *Impulse Control*.

Those SMEs classified as possessing the *Learning* style category of “Dreamer” have a greater chance of displaying performance bias then SMEs who are more organized, industrious, and practical.⁸² Participants who express the *Activity* style category of “Go-Getters” have less of a chance of exhibiting performance bias than SMEs who are more undirected and reserved.⁸³ Finally, SMEs demonstrate the *Impulse Control* style category “Overcontrolled” have less chance of revealing performance bias

⁸¹ The four categories within the *Attitude* personality style are “Free Thinkers” (greater openness and less agreeableness), “Progressive” (greater openness and greater agreeableness), “Result Believers” (less openness and less agreeableness), and “Traditionalists” (less openness and greater agreeableness) [COST 00].

⁸² The four categories within the *Learning* personality style are “Dreamers” (greater openness and less conscientiousness), “Good Students” (greater openness and greater conscientiousness), “Reluctant Scholars” (less openness and less conscientiousness), and “By-the-Bookers” (less openness and greater conscientiousness) [COST 00].

⁸³ The four categories within the *Activity* personality style are “Fun Lovers” (greater extraversion and less conscientiousness), “Go-Getters” (greater extraversion and greater conscientiousness), “Lethargic” (less extraversion and less conscientiousness), and “Plodders” (less extraversion and greater conscientiousness) [COST 00].

than SMEs who are free spirited, weary of controlled environments, or are self-serving.⁸⁴ In fact, zero SMEs with “Overcontrolled” characteristics were amongst those displaying performance biases.

Analyzing the interaction between anchoring bias and the personality demonstrates no statistically significant effect. Further analysis of the personality terms generated from the NEO-FFI scores indicates there is only one personality term demonstrating statistically significant effect on anchoring bias, *Activity*. Thus, with regard to personality and anchoring bias, there are no inferences. Participants who are “Lethargic” are more likely to express anchoring bias than those who are energetic or goal-directed.

The statistically significant effect demonstrated between personality terms and three of the four biases addressed in this research warrant an investigation of the effect of personalities on bias SMEs. As with confirmation bias, an analysis of the interaction between the combined bias and these personalities demonstrated a statistically significant effect from the *Interest* category “Introspectors”.

Analysis of effects based on other personality styles, identifies three additional styles demonstrating an effect on bias, *Attitudes*, *Character*, and *Interactions*. Participants who display bias also show more of the *Interactions* characteristic of “Leaders”.⁸⁵ This indicates individuals who are self-assured and engaging demonstrate less bias. Participants displaying bias exhibit few facets of the *Character* style trait “Self-Promoters” than individuals who are more pragmatic or unconventional prove more bias.⁸⁶ Finally, SMEs who present many of the features of the *Attitude* style traits of a

⁸⁴ The four categories within the *Impulse Control* personality style are “Undercontrolled” (greater neuroticism and less conscientiousness), “Overcontrolled” (greater neuroticism and greater conscientiousness), “Relaxed” (less neuroticism and less conscientiousness), and “Directed” (less neuroticism and greater conscientiousness) [COST 00].

⁸⁵ The four categories within the *Interactions* personality style are “Leaders” (greater extraversion and less agreeableness), “Welcomers” (greater extraversion and greater agreeableness), “Competitors” (less extraversion and less agreeableness), and “Unassuming” (less extraversion and greater agreeableness) [COST 00].

⁸⁶ The four categories within the *Character* personality style are “Well-Intentioned” (greater agreeableness and less conscientiousness), “Effective Altruists” (greater agreeableness and greater conscientiousness), “Undistinguished” (less agreeableness and less conscientiousness), and “Self-Promoters” (less agreeableness and greater conscientiousness) [COST 00].

“Free-Thinkers” or “Traditionalists” exhibit bias. This shows those individuals who are more open and less agreeable or who are less open and more agreeable tend towards bias responses.

Although SME’s who demonstrated certain personality traits, SMEs with these personality traits did not show statistically significant probability of having a bias. Therefore, there personality traits cannot be used to identify SMEs with a likely hood of demonstrating bias.

VI. CONCLUSIONS, SUMMARY, AND RESEARCH AGENDA

Increasing reliance on the use of virtual and constructive models to provide military leaders with information on which to base decisions for development of new weapon systems, reorganizing force structures, and developing tactics, emphasizes the need for more advanced human behavior representation models. With this increased need for higher-fidelity HBR models comes the matter of validation. This has proven to be a difficult and expensive process for the M&S community. To assist the community, this research provides insights into issues regarding the usage of subject matter experts in the face validation of human behavior representation models via overt behaviors. The results of this research are based on data collected as part of an effort to validate a behavioral model utilizing a MAS representation in an entity level, ground combat simulation.

A means to enhance the face validation process for HBR models was used. It identified issues related to consistency and accuracy, effects based on bias and personality, and a means to mitigate these effects. The validation process required a referent with which to compare the model results, a sequence of military scenarios to exercise the model, and a series of sensitivity tests to indicate variance in SME responses. The remainder of this chapter is divided into four sections. The first addresses the conclusions of this research. The second provides a summary of results. Third is a recommended list of procedures for conducting face validation. The final section provides a research agenda intended to further improve face validation procedures for HBR models.

A. CONCLUSIONS

This research identified and/or statistically illustrates nine fundamental conclusions with respect to the use of SMEs in the conduct of the model assessment phase of face validation.⁸⁷ These conclusions are:

⁸⁷ A cautionary footnote: The findings of this research describe the performance of the specialized class of SMEs trained in the field of ground combat and any attempt to generalize beyond this research to other categories of military or civilian personnel must be done with care.

- (1) There is a statistically significant effect based on the scale used to assess performance that can increase or decrease scores for inter-SME consistency and intra-SME consistency, consistency impact, accuracy, and accuracy impact.
- (2) The use of MTP assessment worksheets for assessing simulated human behaviors is a valid as using the worksheets for assessing human performance.
- (3) There is a statistically significant effect based on SME performance bias that can increase or decrease inter-SME consistency and intra-SME accuracy impact.
- (4) There is a statistically significant effect between SMEs demonstrating specific types of bias and exhibiting or failing to exhibit personality categories or styles identified by the NEO-FFI personality inventory.
- (5) For a given subtask, task, scenario, or overall assessment there is a lack of inter-SME consistency.
- (6) For a given SME-subtasks-task combination, there is a lack of intra-SME consistency in the way a SME derives ratings, meaning the given task score is inconsistent with the derived score for the subtasks, which also holds true for SME-tasks-scenario and SME-scenarios-overall combinations. Notably, however, there was no apparent tendency by SMEs to fixate on a specific subtask, task, or scenario, which would have allow the elicitation of a general weighting factor for the subtask, task, or scenario and, in turn, used to explain the lack of consistency between the level score and sublevel assessments.
- (7) For a given SME-subtasks-task combination, there is a practical effect based on the lack of intra-SME consistency, meaning the impact of a given subtask-task pairing inconsistency score's causes a change in the task assessment from Go to No-Go, Go to Unknown, No-Go to Go, No-Go to Unknown, Unknown to Go, or Unknown to No-Go, which holds true for SME task-scenario and scenarios-overall combinations.
- (8) SMEs, on average, are not accurate with respect to a consistent baseline assessment and this inaccuracy is either increased or decreased depending on the scale used.
- (9) In general, for a given level, there is a lack of intra-SME accuracy, with respect to a consistent baseline assessment, causing a change in the task assessment from Go to No-Go, Go to Unknown, No-Go to Go, No-Go to Unknown, Unknown to Go, or Unknown to No-Go, which holds true for all levels of assessment.

B. SUMMARY

Reference conclusion 1, refers to the significant ANOCAT results in the comparison of the absolute value of the differences in SME scores for consistency, consistency impact, accuracy, and accuracy impact, indicate scale can mitigate effects on

these scores. ANOCAT results based on scale and simulation belief indicate Scale 2 (Go/No-Go Scale) provides greater consistency for both inter- and intra-SME scores than the other two scales.

Reference conclusion 2, refers to the ANOCAT results in the comparison of the number of participants displaying performance, anchoring, confirmation, and contrast bias, indicates simulation belief demonstrated no statistically significant effect on these numbers.

Reference conclusion 3, refers to the significant ANOCAT results in the comparison of the absolute value of the differences in SME scores for consistency, consistency impact, accuracy, and accuracy impact, indicate SME biases have an effect on inter- and intra-SME consistency. Table 14 (page 94) and Figure 29 (page 95) are examples of the positive effect on inter-SME consistency when scores for SMEs identified as displaying one or more of the biases are removed from the sample data.

Reference conclusion 4, refers to the stepwise logistical regression and categorical analysis results. Personnel demonstrating one or more of the four bias types studied in this research showed a propensity toward certain categories of *Attitude*, *Character*, *Interactions*, and *Interest* personality styles. In general, a data review indicates SME personality, as characterized by the NEO-FFI, exhibits no predictive capability ($R^2 \leq .0534$) on the inter-/intra-SME consistency and accuracy of assessment. Thus, choosing SMEs based on the five NEO-FFI personality factors or ten personality styles may not allow one to statistically impact SME consistency or accuracy of assessment. However, the results indicate an area validating agents can begin looking for issues with consistency or accuracy of results.

Reference conclusion 5, the process of inter-SME consistency is demonstrated (Figure 30, page 96 and Figure 31, page 97) by independently displaying all SME responses for each subtask, task, scenario, and overall assessment value. Based on the Prob>Chi Square values from an ANOCAT where assessment scale (7-Point Likert, 5-Point Likert, or Go/No-Go) and simulation belief are factors, SMEs' assessment scores for specific subtask, task, scenario, and overall assessment values are distributed across

all possible assessment values (Table 15, page 98). This inconsistency precludes an accurate assessment of the face validity of the simulation. Based on these results, the scale used can mitigate the degree of inconsistency across SMEs, providing inter-SME results that are more consistent. The ANOCAT results indicate simulation belief does not have an effect on inter-SME inconsistency results. According to these results, the scale used can mitigate the lack consistency across SMEs responses.

Reference conclusion 6, a SME's intra-SME consistency score can be calculated using the difference between the level assessment (e.g. task, scenario, and overall) and the mean of the sublevel values (e.g. subtask, task, and scenario, respectively). Based on the Prob>Chi Square values from an ANOCAT where assessment scale (7-Point Likert, 5-Point Likert, or Go/No-Go) and simulation belief are factors, SMEs' assessment scores for a specific level do not reflect the scores derived from the sublevel assessments (Table 16, page 100). This inconsistency precludes an accurate assessment of the face validity of the simulation. According to these results, the scale used can mitigate the degree of internal inconsistency, providing intra-SME results that are more consistent (Figure 33, page 102). The ANOCAT results indicate simulation belief does not have an effect on intra-SME inconsistency results. According to these results, the scale used can mitigate the lack of individual SME consistency.

Reference conclusion 7, SME internal consistency impact scores are based on the number of level assessments (e.g., task, scenario, and overall) which flip-flop their binary Go/No-Go assessment based on the difference between the level assessment (e.g. task, scenario, and overall) and the average of the sum of the sublevel values (e.g. subtask, task, and scenario, respectively) per SME. This calculation provides the impact of a SME's intra-SME inconsistency based on a change in level rankings (e.g. Go to No-Go, Go to Unknown, No-Go to Go, No-Go to Unknown, Unknown to Go, or Unknown to No-Go). Ascertained from the Prob>Chi Square values from an ANOCAT with the factors of scale and simulation belief, the effect on the intra-SME scores is inconsistent with an effect based on scale. The ANOCAT results indicate simulation belief does not have an effect on intra-SME consistency impact results (Table 17, page 104). Impact inconsistency provides insight into the bearing the SMEs' inconsistencies have on the

overall assessment of the simulation by demonstrating how intra-SME inconsistency results in a final assessment, which differs, from the summation of the sublevel assessments. Examining the results based on the scale used (Figure 34, page 103), it is apparent that one can mitigate the practical effect of the intra-SME inconsistency for a SME and reduce the overall inconsistency between the sublevel assessments and the overall assessment for a SME.

Reference conclusion 8, a SME's accuracy scores are the difference between a SME's raw assessments and the assessment key, which has consistent baseline assessment values. Based on the Prob>Chi Square values from an ANOCAT with the factors scale and simulation belief, SMEs are inaccurate in their assessment with scale having a statistically significant effect on the magnitude of the inaccuracy (Table 17, page 105). Inaccurate ratings prevent a coherent assessment of the face validity of the simulation. According to these results, the scale used can mitigate the degree of inaccuracy.

Reference conclusion 9, SME internal accuracy impact scores are based on the number of assessments (e.g. scenario, task, scenario, and overall) which flip-flop their binary Go/No-Go assessment based on the difference from the assessment key score for each question. This calculation provides the impact of a SME's intra-SME inaccuracy based on a change in question rankings (e.g. Go to No-Go, Go to Unknown, No-Go to Go, No-Go to Unknown, Unknown to Go, or Unknown to No-Go). Ascertained from the Prob>Chi Square values from an ANOCAT with the factors of scale and simulation belief, the effect on the intra-SME scores is inaccurate with an effect based on scale. The ANOCAT results indicate simulation belief has an effect on intra-SME accuracy impact scores at the subtask and task level (Table 18, page 108). Impact inaccuracy measures the practical effect SME inaccuracies have on the SME's overall assessment of the simulation. Displaying the results based on scale (Figure 36, page 109), it is apparent that the scale used can mitigate the practical effect of the intra-SME inaccuracy reducing the overall inaccuracy between SME assessments and a consistent baseline assessment.

C. FACE VALIDATION PROCESS RECOMMENDATIONS

Through the process of conducting research, several ideas have been fostered which would assist in the validation of HBR model implementations. These will be addressed in the context of the validation procedures outlined in Chapter 2.

A Validation Plan must be created to outline the purpose of the validation process, the measures of effectiveness, the measures of performance, the tasks to be assessed, data collection techniques, the training process for SMEs, the techniques for analysis the assessment data, the measures for assessing valid and invalid performance, and the means for deconflicting SME assessments. The plan should also outline the responsibilities for each participant to include who will collect and validate the referents to be used in the validation of the model.

Once the purpose of the model has been identified and the plan completed, referents are collected for use in the validation of the model. For HBR model implementations, this should be based on similar standards to those used for assessing human performance of the real-world tasks. Examples of baseline overt behaviors can be seen in the Mission Training Plans used by the military to describe the tasks and subtasks soldiers must perform in order to receive a status as trained. Textural examples of how to properly perform these tasks can be found in military training manuals. However, one must also provide written examples of improper and ambiguous performance of these tasks to help set boundary conditions for assessing the performance of the model.⁸⁸ The written examples must be supplemented with visual examples for both the real-world and simulated environments so SMEs can correlate the two and better understand what they should be looking for (i.e. increase consistency between raters).

Using mission training plans for collecting referents provides viable criteria for the assessment of obviously proper and improper overt behaviors; however, other data will be needed for the assessment of ambiguous overt behaviors. To assess these behaviors, context of the situation and the situational understanding of the executing

⁸⁸ Ambiguous performance is performance which is neither obviously proper nor improper, but can only be classified based on the context on which it is or is not performed (e.g. if a task says the soldier should use his night vision devices, this may not be appropriate if it is day time, the battlefield is illuminated, or the night vision devices are inoperable).

entity must be understood. This requires collection of referents based on cognitive process. These referents can be gathered using the CTA model to interview SMEs used for the referent collection phase. After referents are collected, they must be validated by a second set of SMEs and then farmed by the validating agent to identify those pertinent to the validation of the HBR model implantation.

After validated referents have been identified for use in the validation of the model, assessment worksheets and scenarios are developed to provide a means to collect data. Worksheets can be automated to allow for rapid importation of data and to assist SMEs in maintaining consistency in their assessments. If any level has more than five sublevel assessments which make up its assessment (e.g. task A has 15 subtasks or scenario B has 6 tasks), then a Go/No-Go scale or computerized assessment system should be used to ensure SMEs do not run into problems with maintaining consistence between the level assessment and its sublevel assessments. Aggregate level assessments should use a 7-Point Likert Scale to provide SMEs with the ability to reflect the degree to which a model behavior is in compliance. Worksheets must always provide room for comments for SMEs to express views not reflected in the assessment questions by means of the assessment scale.

A Pilot Study should always be conducted to work out any problems with the scenarios, data collection procedures, and assessment work sheets. The pilot study must follow the same phases as the main study.

SMEs used in the process of collecting referents, validating referents, or validating the model implementation should take a personality test in order to identify if they may be overly critical or lenient. SME personalities may not have an effect on the as results of the validation, however, if conflicting results are reached, the personality test may indicate which SME data maybe a cause of undue variance in the process.

All SMEs used in the model validation process must under go training to ensure they understand the assessment procedures. This training should include:

- Familiarization with the general assessment process;
- Familiarization with the use of the assessment worksheets;

- Familiarization with the importance of each component of the data provided by the SMEs;
- Practice on the use of the worksheets;
- Familiarization with the model functions and displays;
- Focus of the validation process;
- Written and visual examples of real-world and model performance which demonstrate proper, improper and ambiguous performance of the tasks;
- Assessing of SMEs proficiency with the assessment procedures and materials to assess scripted scenarios which demonstrate obviously poor and obviously improper performance of tasks; and
- Retraining and reassessment of SMEs as necessary.

Assessing SME performance is necessary to ensure they are following the proscribed procedures. It also provides an opportunity to identify if SMEs are demonstrating patterns of bias which create issues with the SMEs consistency and accuracy. If SMEs are demonstrating bias and/or problems with consistency or accuracy, SMEs can be retrained and reassessed to mitigate these issues prior to the validation of the model. If SMEs cannot perform the validation procedures without demonstrating inconsistencies or inaccuracy which adversely impact the validation process, the validating agent needs to determine if a new set of SMEs is needed, if the procedures should be modified, or if the inconsistencies and inaccuracy can be effectively be addressed in the analysis of the data.

During the data collection, SMEs must be continually monitored to ensure they are following the proscribed procedures and to assist them with running the scenarios and model. Care must be taken to ensure SMEs maintain focus on the tasks to be assessed. Additional assessment of the model may be performed only if it does not detract from the primary focus of the validation process.

At the end of the validation process, SMEs should undergo a one-on-one debriefing. The debriefing provides a chance to thank them for their efforts, reiterate the purpose of the work they just performed and to provide them with the opportunity to

make any additional comments concerning the validity of the model and the effectiveness of the procedures. The validation agent can also use this as an opportunity to review and deconflict any misunderstandings with the SME's responses.

Analysis of SME quantitative responses should begin with an examination of SME consistency. Areas which lack consistency helps identify tasks where qualitative data or further SME interviews may provide clarification. This must be done as soon as possible before the SME is too far removed from the process to remember any specifics. Consistency is one means to help identify if the process is potentially under control and sufficient for provide viable results.

If SME responses are internally consistent, the next analysis is to determine consistency amongst SMEs for each response question. For this consistency check, one should aggregate answers into Go/No-Go groupings. This consistency is concerned with general proficiency of each task and not on the degree of proficiency. If there is a lack of inter-SME consistency, then the validating agent must examine the question and qualitative responses to determine which SME responses are most appropriate. Inconsistency must be noted in the validation report and the validating agent should attempt to gain clarification from SMEs who provided outlier responses.

With inter-SME consistency established, the validity of each subtask, task, and scenario can be analyzed. The results of these phases of analysis are then integrated into a report identify the strengths and limitations of the model, the boundaries under which it can perform appropriately, and the overall validity of the model implementation.

D. TOWARDS A RESEARCH AGENDA

To further investigate the intersection of the overlapping ovals of the methodology, this section outlines additional research areas designed to enhance face validation procedures for human behavior representation models. The fundamental issue is not whether the M&S and Psychology Communities need HBR models or that face validation is necessary. The issues are how to build better HBR models and how to conduct validation in a more consistent, accurate, and cost effective manner.

The goal is to create a library of HBR model implementations which have undergone the VV&A process. The library would provide access to accreditation documents which would provide a listing of known model limitations and boundary conditions for its proper use. The known performance characteristics would allow users to take a model off the shelf, integrate it into a simulation and, if the proper boundary conditions are adhered to, the model should perform properly.

1. Issues with Human Behavior Representation Models

As stated in the first chapter, one factor making the validation of HBR model implementations difficult is the nonlinear relationship between inputs and output behaviors. To resolve this issue we must investigate decision-making models and ensure they explanation their cognitive processes in a manner understandable by SMEs and validating agents. This requires the development of new cognitive architectures capable of explaining what they understood as the situation, what options they considered, and the reason they took certain actions.

This requires an understanding of the highly desirable properties of human behavior representation models, without modifications or compromises imposed by cost or other constraints. In other words, the ultimate HBR model implementation for use in DoD research and training.

First, models must exhibit behaviors within the commonly accepted limitations of human behavior. To establish these boundaries, professionals in the field of psychology, biology, and domain specialists must identify and substantiate universally accepted constraints for human behavior. Models must not permit superhuman behavioral performance; however, they should demonstrate fringe behavior with the probability of such behavior falling within ‘realistic’ limits. Not every soldier can run the forty-yard dash in five seconds, and not every leader can assimilate information provided from all available resources and come to the ‘approved’ solution.

Human behavior models must be adequate for the environment they are designed to operate in and should be able to adapt to realistic changes in that environment. Examples of these changes include weather effects, dynamic terrain effects, mission

constraints, commander's intent, enemy actions, etc. Modified behaviors, due to changes in the environment, must be 'realistic' for the current conditions and a logical continuation from previous conditions.

As with humans, an HBR model should help SMEs determine if the model's actions/behaviors are reasonable. Models provide a means for questioning behaviors. Questioning and understanding the situational awareness of the model and its logic process(es) allows SMEs to better determine if the model's situational awareness is adequate, its list of options complete, its logic process accurate, and its overt actions are viable. Identify shortcomings in the logic process or database of possible behaviors can provide a more meaningful assessment of the model's actions.

As stated above, professional expertise is needed to conduct this research. For the areas considered in this research are very complex and highly interdisciplinary. In order to integrate relevant knowledge and the planning of future research in a fashion useful to the pursuit of the declared goal, the M&S community must hold periodic workshops. Many DoD M&S organizations currently conduct workshops with the general focus of VV&A.⁸⁹ However, additional workshops, which integrate the fields of psychology, HBR, combat M&S, and operations research, can help focus efforts and promote interaction between these communities in the area of human behavior representation models. These workshops should involve but not be limited to representatives of the communities associated with virtual environments, combat M&S, operations research, psychology, training, personnel evaluation, and specific application domains. Those managing the M&S policy must continue to give particular thought to the prioritization of the various possible application domains. In addition, policy managers must continue to carefully judge how best to organize and oversee the research. In particular, special consideration must be given to the formation and funding of cross-institutional consortia to deal with the development of HBR models in manner which enhances the abilities to perform VV&A.

⁸⁹ *Foundation for Modeling and Simulation (M&S) Verification and Validation (V&V) in the 21st Century* (Foundations '02), the Military Operations Research Society's (MORS) *Test & Evaluation, Modeling and Simulation and VV&A: Quantifying the Relationship Between Testing and Simulation* workshop, and the NAVMSMO VV&A Technical Working Group (TWG) workshops are examples of efforts to conduct general and domain specific workshops on the VV&A process.

2. Validation Issues

A second factor making HBR model implementations validation difficult is the variability in evaluations based on the consistency and accuracy of SMEs. To resolve this one must address numerous issues: the standards on which assessments are made, the use of subject matter experts, and the procedures used.

a. Referent

An effort is currently underway by Program Executive Office for Simulation, Training, and Instrumentation (PEO STRI) to develop referent for the behavior module of OneSAF based on the programs knowledge acquisition process. Although this is an admirable effort to quantify aspects of human behavior allowing model developers to codify behaviors and provide a benchmark for validation, it does not address the cognitive aspects of the human behavior selection process. One methodology for capture cognitive information is CTA.

Once dissected in a quantitative manner, model developers must texturally described overt and cognitive behaviors in a manner in which coders and validating agents can easily understand. Validating agents then need to translate this information into a set of criteria for use in assessing model performance. Agents can use this information to provide textural and visual examples of adequate, questionable, and inadequate behaviors.

The development of viable referent, assessment worksheets, and examples (for training programs) is a time consuming and costly endeavor. There are varying categories and sources of referent each with its own intrinsic costs. We must conduct studies to demonstrate the trade offs between the cost of collecting, mining, and validating different categories and quantities of referent and the consistency, accuracy, completeness, and usefulness of the ensuing model validation results.

b. Subject Matter Experts for Assessment Process

Although there are many issues with the use of SMEs, computability theory indicates we must still use SMEs in order to assess models of human behavior. Since human behavior is non-deterministic, one cannot write an algorithm to assess if a deterministic program, which is replicating non-deterministic behavior, is performing

correctly; heuristics apply but are not absolute. Thus, since the use of SMEs is necessary for the validation of HBR models, additional research is required to address issues with categorizing, training, certifying, and supervising SMEs.

The Defense Modeling and Simulation Office, Pace, and Klein have described criteria for SME selection and certification. However, criteria such as years of experience beg issues. Requiring an individual to have ten years or ten thousand hours of experience in a specialty area to warrant certification as an expert can produce a SME population that has depth and diverse experience. Such a time requirement may result in their removal from duty positions, due to promotion, to place SMEs out of contact with state of the art systems and their implementation rendering SMEs virtually obsolete for validating human behavior representation models for entry-level duty positions.

The M&S Community must investigate this and other issues. We must conduct research to measure the level of training versus experience required to possess the sufficient skills required to be a SME. Other research is required to determine how combat, peace enforcement, peacekeeping, and other real-world experiences equate to training/simulated experience. How far removed, can a SME be from the discipline he is assessing the behaviors of and still be a viable for assessing behaviors in that domain?

This research utilized NEO-FFI to categorize individual personalities for use in determining if personality has a significant effect on SME assessment. Although certain personality traits demonstrated a statistically significant effect on assessment responses and consistency and accuracy scores, their predictive capability was nearly non-existent. However, the population did not provide a complete cross section of personality types as categorized by the NEO-FFI. We must conduct additional research explicitly designed to determine if the NEO-FFI or other personality tests provide personality classifications that predict SME responses. If the studies identify predictive tendencies towards consistency and accuracy based on personality characteristics, we can exploit these characteristics for the selection of SMEs for use in performing face validation.

Pace and Sheehan at Foundations '02 and DMSO VV&A TWG discusses the need for training SMEs in the validation process. As with experience, training begs

numerous questions. What specific training must an individual undergo to receive certification as a SME for HBR models? What criteria must a SME meet to receive certification as a SME for HBR models? What knowledge and skills must an individual demonstrate to receive certification as a SME for HBR models? Does this level of training make a SME adequately prepared to validate an HBR model? Is a centralized training program required to ensure base knowledge of all SMEs used in the validation process? What should a training program cover? What is the most effective training program? How often must a SME undergo retraining to maintain his SME certification? These are just a few of the questions studies must address to ensure SMEs are adequately prepared to perform the task of HBR model validation. The M&S community must investigate these and other training issues to ensure completeness of programs and uniformity of standards.

Literature reviews reveal limited to no published research in the area of assessing individuals who are evaluating the performance of real-world or simulated performance. However, we do not need to wait for the next generation human behavior model implementations with architectures that can explain themselves. We can investigate this issue leveraging human performance evaluation techniques. Through the study of experts who are assessing human performance, we can gain insight into what questions should be asked and how they should be answered. This would help drive the development of the new HBR architectures.

This research classified four types of SME bias and identified patterns that we can use to recognition them. The research also showed the effects on consistency and accuracy when SMEs demonstrating these biases have their data removed from the sample pool. Next, the research identified possible modification methods that have shown to mitigate bias in the assessment of human performance in the work place: computerized validation tool, grounding scales, weighting the importance of criteria, etc. The next step is to conduct further research to determine if these modifications to the assessment scales have a statistically significant effect on reducing the presence of bias and increasing accuracy and consistency. Although this research provides statically significant evidence of the existence and effect of SME bias, we must ensure the limited number SMEs

available to the validating agent are capable of performing an unbiased assessment of model performance through the proposed modifications to the validation process.

Each technique for mitigating bias requires additional resources. Further research must capture the cost and the degree of process enhancement. This will help to develop a cost function for end users to measure tradeoffs and assess the risk of not implementing enhanced procedures.

c. Validation Procedures

Development of procedures that constrain subjectivity is essential to maintain process control and providing consistent and accurate results from which to translate into an accreditation document. To facilitate validation, one must have a consensus on how to express overall model validation; how to represent model process and query them for clarification; and what measures to use to collect consistent and accurate data for use in the validation process.

Harmon et al. and Goerger propose two discrete but related theories to enhance the general validation process for HBR, which modify the face of the final validation product. Instead of valid or invalid, Harmon et al., recommend categories of validation, while Goerger suggests a sliding scale of validity [HARM 03] [GOER 03]. Conducting studies to determine the feasibility and usefulness of these theories is required.

Another aspect of the face validation process, which must be addressed, is the manner in which the model presents data to SMEs. One might enhance the validation process by modifying the manner in which models display their behaviors. Due to their number of elements and the scope of many analytical models, they routinely present behaviors on a 2D map display or in textural records. Presenting information using 3D models in a stealth view may provide additional information to SMEs. 3D models allow SMEs to observe model behaviors in the same manner that evaluators follow soldiers through the environment in training exercises. This could potential clarify model behaviors in a manner which 2D displays are incapable. For example, if a SME sees an icon representing a soldier moving through an urban environment stop along the edge of building just short of a window for two to three minutes he may not be able to tell the extent of behaviors the icon is executing. When displayed in a 3D environment, the SME

may see a disoriented entity checking its map, an entity stopping to fix his equipment, or an entity attempting to crawl through the wall because it cannot identify the window location. Without the information on the posture and activity of the entity, the SME is left to his own imagination to the status of the entity. We need to conduct research in the effectiveness of 2D and 3D displays in providing information to SMEs to determine the level of information the displays provide, their impact on assessment scores, and their cost effectiveness ration.

A corollary effort is the ability to query model implementations for information. This is similar to an after-action review or interview of the model. To enhance a SMEs ability to understand the procedural aspects of the model's overt actions it would be useful to question a model about its situational awareness, possible courses of action, and thought process. A model's ability to provide SMEs with such information would give them a better understanding of why an HBR model implementation performed certain actions. This enhances their ability to make a comprehensive assessment of the model. Two things must occur to assist the investigation into types of and need for information for face validation. First, an HBR model implementation must be built that would allow a query of the model. Second is experimentation to test the difference in SME assessments of HBR model implementations which differ only in their ability to provide insight about why it performed certain actions. These tasks help to identify information tracking features required in an HBR model to assist in its verification and validation effort.

An alternative to building an HBR model implantation SMEs could query is to conduct a Turing Test type experiment allowing SMEs to question an HBR model implementation using a text window. Research personnel would respond to queries using a text window to simulate the model's hypothetical ability to provide requested information. This research could identify the difference in SME assessments of HBR overt actions pending their ability to query the model. The research would also gain incite into the type and amount of information SMEs need to assist in making an accurate and consistent validation of an HBR model implementation.

Finally, as stated earlier, standardization and enhancements to validation procedures through the grounding and weighting of assessment criteria are possible solutions to reducing SME bias and enhancements to consistency and accuracy in the validation of HBR models. We must conduct additional studies focusing on the use of these techniques to ensure their feasibility and effect.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A. REFERENT FOR HUMAN BEHAVIOR REPRESENTATION MODELS

The referent used for the assessment of the human behaviors is based on data from *FM 7-8: Infantry Rifle Platoon and Squad*, 2001 and *ARTEP 7-8-MTP: Mission Training Plan for the Infantry Rifle Platoon and Squad*, 2001 [DEPA 01g] [ARTP 01]. This appendix provides a sample of one of the evaluation forms from this manual. The assessment sheets developed for this research come from a distillation of this document (Appendix K. Assessment Worksheets).

07-3-1406

NOTICE:

This document is generated from relational data submitted by the proponent. The complete, authenticated document, when available, may be downloaded from the "Official Departmental Publications" side of the RDL.

Questions relating to information displayed should be addressed to the proponent school.

TASK: React to Snipers (Infantry/Reconnaissance Platoon/Squad) (07-3-1406)

([FM 21-60](#)) (FM 24-35) (FM 24-35-1) (FM 7-4 (3-21.94)) (FM 7-5 (3-21.9)) ([FM 7-7](#)) ([FM 7-7J](#)) ([FM 7-8](#)) ([FM 7-85](#)) ([FM 7-92](#)) (FM 90-10(HTF)) ([FM 90-10-1](#))

ITERATION	1	2	3	4	5	M (circle)
TRAINING STATUS	T	P	U			(circle)

CONDITION: The platoon is conducting operations as part of a larger force and receives fire from an enemy sniper. The platoon must react immediately for their protection. All necessary personnel and equipment are available. The platoon has communications with higher, adjacent, and subordinate elements. The platoon has been provided guidance on the rules of engagement (ROE) and or rules of interaction (ROI). Coalition forces and noncombatants may be present in the operational environment. Some iterations of this task should be conducted during limited visibility conditions.

Some iterations of this task should be performed in MOPP4.

TASK STANDARD: The platoon reacts to the sniper in accordance with (IAW) tactical standing operating procedures (TSOP), the order, and or commander's guidance. The platoon correctly locates and then bypasses, eliminates, or forces the withdrawal of the enemy sniper while disengaging the element in the kill zone. The platoon complies with the ROE and or ROI.

TASK STEPS and PERFORMANCE MEASURES	GO	NO GO
<p>1. Platoon conducts actions on contact (sniper fire).</p> <ul style="list-style-type: none"> a. Returns fire immediately to destroy or suppress the enemy. b. Deploys to covered and concealed positions, if available. c. Utilizes indirect fire assets, if available. d. Activates on board self-protection measures as appropriate. e. Conducts battle drills, as necessary. f. Maintains visual contact with the enemy while continuing to develop the situation through reconnaissance or surveillance. g. Maintains cross talk with all platoon elements using FBCB2, FM, or other tactical means. <p>2. Platoon reacts to enemy sniper fire.</p> <ul style="list-style-type: none"> a. Reports contact to higher headquarters using FBCB2, FM, or other tactical means. b. Bypasses the sniper. <ul style="list-style-type: none"> (1) The platoon uses smoke to obscure the enemy snipers view. (2) The platoon uses available fires to suppress the sniper. (3) The platoon maneuvers to break contact with the sniper. Note. The platoon leader may choose to call for indirect fire on the sniper position. 		

TASK STEPS and PERFORMANCE MEASURES	GO	NO GO
<p>c. Eliminates the sniper.</p> <p>(1) Complies with ROE and or ROI.</p> <p>(2) The platoon uses smoke to obscure the enemy snipers view.</p> <p>(3) The platoon uses available firepower to suppress and fix the sniper.</p> <p>(4) The platoon maneuvers to close with the sniper and eliminate or force him to withdraw.</p> <p>3. Platoon consolidates and reorganizes as necessary.</p> <p>4. Platoon treats and evacuates casualties as necessary.</p> <p>5. Platoon secures enemy prisoners of war (EPW), if applicable.</p> <p>6. Platoon processes captured documents and or equipment, if applicable.</p> <p>*7. Platoon leader reports to higher headquarters as required using FBCB2, FM, or other tactical means.</p> <p>8. Platoon continues operations as directed.</p>		
<p>NOTE * Indicates a leader task. NOTE + Indicates a critical task.</p>		

TASK PERFORMANCE SUMMARY BLOCK							
ITERATION	1	2	3	4	5	M	TOTAL
TOTAL TASK STEPS & PERFORMANCE MEASURES EVALUATED							
TOTAL TASK STEPS & PERFORMANCE MEASURES "GO"							

SUPPORTING COLLECTIVE TASKS

- 07-3-1009 Conduct a Deliberate Attack (Infantry Platoon/Squad)
- 07-3-1045 Conduct a Bypass (Infantry/Reconnaissance Platoon/Squad)
- 07-3-1072 Conduct a Disengagement (Infantry/Reconnaissance Platoon/Squad)
- 07-3-1252 Conduct Overwatch and or Support by Fire (Antiarmor/Infantry Platoon/Squad)
- 07-3-1270 Conduct Tactical Movement (Mounted or Dismounted)
(Antiarmor/Infantry/Mortar/Reconnaissance Platoon/Squad)
- 07-3-1279 Conduct Tactical Movement in a Built-up Area
(Antiarmor/Infantry/Reconnaissance Platoon/Squad)
- 07-3-1432 Take Action on Contact (Infantry/Mortar/Reconnaissance Platoon/Squad)
- 07-3-2054 Report Tactical Information (Infantry/Mortar/Reconnaissance Platoon/Squad)
- 07-3-4009 Handle Enemy Prisoners of War (Infantry/Mortar/Reconnaissance Platoon/Squad)
- 07-3-4027 Process Captured Documents and Equipment
(Infantry/Mortar/Reconnaissance Platoon/Squad)
- 07-3-4045 Treat and Evacuate Casualties (Infantry/Mortar/Reconnaissance Platoon/Squad)
- 07-3-5009 Conduct Consolidation and Reorganization (Infantry/Reconnaissance Platoon/Squad)
- 07-3-5036 Conduct Troop-leading Procedures (Infantry/Mortar/Reconnaissance Platoon/Squad)
- 07-3-6027 Maintain Operations Security (Infantry/Mortar/Reconnaissance Platoon/Squad)

OPFOR TASKS AND STANDARDS

- 07-OPFOR-0017 Maintain Operations Security (Infantry/Mortar/Reconnaissance Platoon/Squad)

APPENDIX B. EXPERIMENTAL PROCEDURES

The following are the experimental procedures used for the different phases of the study. The first four phases (In Brief/Consent Form, Assessment Procedure Familiarization, Model Familiarization, and Practical Exercise) of the study were the same for each participant in Study #1. The initial four phases for Study # 2 differed only in the assessment scale taught to the participants (5-Point Liker or Go/NoGo). During the Assessment Phase for Study #1, groups differed only in whether they were told a specific scenario was live or constructive in its generation. During the Assessment Phase for Study #2, groups differed in whether they were told a specific scenario was live or constructive in its generation and in the assessment scale they used (5 Point Liker or Go/NoGo). All participants undergo a debriefing at the conclusion of the assessment phase to hand out their NEO-FFI results, a debriefing questionnaire, and a one-page description of the experiment with points of contact information.

(1) In Brief/Consent Form

- (a) Time – 15 Min
- (b) Location – Study Room
- (c) OIC – MAJ Simon R. Goerger
- (d) Materials – Consent Form, Privacy Act Statement, Minimal Risk Consent Form, Participant Demographics Data Form, Participant Roster, NEO Five Factor Inventory, pencils, In-Briefing Script, In-Briefing PPT

(2) Assessment Procedure Familiarization

- (a) Time – 15 Min minimum
- (b) Location – Study Room
- (c) OIC – MAJ Simon R. Goerger
- (d) Materials – Laptop, test model, Assessment Reference Poster, Assessment Procedure Briefing Script, Assessment Form(s), Assessment Procedure PPT

(3) Model Familiarization

- (a) Time –10 Min minimum
- (b) Location – Study Room
- (c) OIC – MAJ Simon R. Goerger
- (d) Materials – Laptop, model, Model Interface Reference Card, Model Familiarization Briefing Script, Model Interface Reference Poster, Interface Familiarization PPT

(4) Practical Exercise

- (a) Time – 20 Min
- (b) Location – Study Room
- (c) OIC – MAJ Simon R. Goerger
- (d) Materials – Laptop, model, Model Interface Reference Poster, Assessment Reference Poster, Blue Pen, Terrain Sketch, Assessment Form(s), Practical Exercise Briefing Script
- (e) Assessment Practical Exercise (10 min minimum)

(5) Assessment

- (a) Warm-up (assessing performance)
 - 1. Time – 10 Min
 - 2. Location – Study Room
 - 3. OIC – MAJ Simon R. Goerger
 - 4. Materials – Laptop, model, Model Interface Reference Poster, Assessment Reference Poster, Blue Pen, Terrain Sketch, Assessment Form(s), Warm-up Briefing Script

(b) Study Group #1 - Control Group (assessing human performance – three scenarios)

1. Time – 45 Min
2. Location – Study Room
3. OIC – MAJ Simon R. Goerger
4. Materials – Laptop, model, Model Interface Reference Poster, Assessment Reference Poster, Blue Pen, Terrain Sketch, Assessment Form(s), Control Group Briefing Script

(c) Study Group #2 (assessing CGF performance – three scenarios)

1. Time – 45 Min
2. Location – Study Room
3. OIC – MAJ Simon R. Goerger
4. Materials – Laptop, model, Model Interface Reference Poster, Assessment Reference Poster, Blue Pen, Terrain Sketch, Assessment Form(s), Study Group #2 Briefing Script

(d) Study Group #3 (assessing two CGF & one human performance scenarios)

1. Time – 45 Min
2. Location – Study Room
3. OIC – MAJ Simon R. Goerger
4. Materials – Laptop, model, Model Interface Reference Poster, Assessment Reference Poster, Blue Pen, Terrain Sketch, Assessment Form(s), Study Group #3 Briefing Script

(e) Study Group #4 (assessing two human & one CGF performance scenarios)

1. Time – 45 Min
2. Location – Study Room

3. OIC – MAJ Simon R. Goerger
4. Materials – Laptop, model, Model Interface Reference Poster, Assessment Reference Poster, Blue Pen, Terrain Sketch, Assessment Form(s), Study Group #4 Briefing Script

(6) Debriefing

- (a) Time – 5 Min
- (b) Location – Study Room
- (c) OIC – MAJ Simon R. Goerger
- (d) Materials – Debriefing Script, Debriefing Questionnaire, Debrief Statement, NEO-FFI Report, and Americas Army CD

APPENDIX C. PARTICIPANT TASKS

Appendix C is a description of the tasks participants were asked to perform in order to assess the performance of the behaviors they viewed through the MANA interface. Participants assessed three scenarios. All scenarios are executed in an urban environment (McKenna MOUT Site, Fort Benning, GA) by squad size elements. Two scenarios are offensive scenarios and one is a defensive scenario. Participants evaluate an offensive scenario, the defensive scenario, and then the final offensive scenario. After each scenario, participants complete a scenario summary assessment. At the end of the third scenario, participants complete an overall assessment of the performance of the model or squad they assessed during the three scenarios.

1. ATTACK URBAN AREA #1

Task 1. Evaluate Deploy/Conduct Maneuver

- (a) Focus: Individual/Squad Tactical Movement (Dismounted)
- (b) Task: Conduct Tactical Movement in a Built-up Area (Infantry Squad) (07-3-1279)
- (c) Measure: In accordance with FM 90-10-1 and ARTEP 7-8-MTP standards

Task 2. Evaluate Deploy/Conduct Maneuver

- (a) Focus: Survivability
- (b) Task: React to Snipers (Infantry Squad) (07-3-1406)
- (c) Measure: In accordance with FM 90-10-1 and ARTEP 7-8-MTP standards

Task 3. Evaluate Deploy/Conduct Maneuver

- (a) Focus: Individual/Squad Tactical Movement (Dismounted)
- (b) Task: Conduct Tactical Movement in a Built-up Area (Infantry Squad) (07-3-1279)
- (c) Measure: In accordance with FM 90-10-1 and ARTEP 7-8-MTP standards

Task 4. Evaluate Deploy/Conduct Maneuver

- (a) Focus: Individual/Squad Tactical Movement (Dismounted)
- (b) Task: Offensive Scenario #1 Assessment
- (c) Measure: In accordance with FM 90-10-1 and ARTEP 7-8-MTP standards

2. DEFEND URBAN AREA

Task 1. Evaluate Deploy/Conduct Maneuver

- (a) Focus: Individual/Squad Tactical Operations (Dismounted)
- (b) Task: Conduct a Strongpoint Defense of a Building (Infantry Squad) (07-3-1162)
- (c) Measure: In accordance with FM 90-10-1 and ARTEP 7-8-MTP standards

3. ATTACK URBAN AREA #2

Task 1. Evaluate Deploy/Conduct Maneuver

- (a) Focus: Individual/Squad Tactical Movement (Dismounted)
- (b) Task: Conduct Tactical Movement in a Built-up Area (Infantry Squad) (07-3-1279)
- (c) Measure: In accordance with FM 90-10-1 and ARTEP 7-8-MTP standards

Task 2. Evaluate Deploy/Conduct Maneuver

- (a) Focus: Survivability
- (b) Task: React to Snipers (Infantry Squad) (07-3-1406)
- (c) Measure: In accordance with FM 90-10-1 and ARTEP 7-8-MTP standards

Task 3. Evaluate Deploy/Conduct Maneuver

- (a) Focus: Individual/Squad Tactical Movement (Dismounted)
- (b) Task: Conduct Tactical Movement in a Built-up Area (Infantry Squad) (07-3-1279)
- (c) Measure: In accordance with FM 90-10-1 and ARTEP 7-8-MTP standards

Task 4. Evaluate Deploy/Conduct Maneuver

- (a) Focus: Individual/Squad Tactical Movement (Dismounted)
- (b) Task: Offensive Scenario #2 Assessment
- (c) Measure: In accordance with FM 90-10-1 and ARTEP 7-8-MTP standards

4. OVERALL ASSESSMENT

Task 1. Evaluate Deploy/Conduct Maneuver

- (a) Focus: Overall Performance
- (b) Task: Summarize overall performance during scenarios
- (c) Measure: In accordance with FM 90-10-1 and ARTEP 7-8-MTP standards

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX D. ASSESSMENT OF PARTICIPANT TASKS

Appendix D. is a description of the nine tasks that provide the focus for this research. The tasks help to identify and address issues with the use of SMEs and the effect of assessment scale on these issues. The nine tasks are based on identifying bias, consistency, and accuracy.

Task 1. Assess model/human performance

(a) **Focus:** Anchoring Bias

(b) **Task:** Identify if a participant embraces an initial hypothesis and maintains this view regardless of incoming facts resulting in overemphasis on the hypothesis and an inappropriately minimal shift from the participant's initial viewpoint.

(c) **Measure:** Anchoring bias exists in a participant when the SME judges the first task and associated subtasks as a "Go" and then after viewing the second task and associated subtasks⁹⁰ judges the remainder of the model performance as "Go" for more than 90% of the assessment questions for which he provides a passing or failing appraisal. The SME's remaining appraisals are made with indifference to evidence to the contrary of the assessment value made. Anchoring bias also exists in a participant when the SME judges the first scenario, associated tasks and subtasks as "No-Go" and then after viewing the second scenario and associated subtasks⁹¹ judges the remainder of the model performance as "No-Go" for more than 90% of the assessment questions for which he provides a passing or failing appraisal. Again, the SME's remaining appraisals are made with indifference to evidence to the contrary of the assessment value made.

⁹⁰ In accordance with doctrine, the squad fails to perform properly the second task and associated subtasks for "React to the Sniper Attack" by losing two personnel without the remainder of the squad reacting to the sniper's attack or the loss of personnel.

⁹¹ In accordance with doctrine, the squad properly performs the second scenario and associated task and subtasks as it successfully defends the building by destroying enemy forces attempting to seize the structure.

Task 2. Assess model/human performance

(a) **Focus:** Contrast Bias

(b) **Task:** Identify if a participant seeks information to contradict his original hypothesis, ignoring or under valuing evidence in support of the hypothesis.

(c) **Measure:** Contrast bias exists in a participant when the SME starts with a negative or positive opinion and after viewing data differing from this initial opinion, the participant negates any further evidence in support of the original hypothesis and assesses the model based on the swing of opinion. In addition, the SME's accuracy data must indicate a shift in his accuracy trend, from harsher to more lenient or more lenient to harsher, as the assessment process proceeds. This shift occurs after the swing in raw score responses.

Task 3. Assess model/human performance

(a) **Focus:** Confirmation Bias

(b) **Task:** Identify if a participant overvalues select pieces of information providing an inconsistent assessment of performance relative to evidence indicating an alternate conclusion.

(c) **Measure:** Confirmation bias exists in a participant when the SME's differences in sublevel mean scores and level responses trend towards no difference in response or show a generally consistent difference in response and the overall response differs from this trend.⁹²

Task 4. Assess model/human performance

(a) **Focus:** Performance Bias

(b) **Task:** Identify if a participant is hampered in his ability to perform face validation by other demands on his time, the availability of data, the ability or desire to comply with specified validation procedures, or the ability of the expert to understand the simulation.

⁹² Note: differences between sublevel mean scores and level responses may mitigate each other with the addition of more assessment responses; this does not indicate confirmation bias.

(c) **Measure:** Performance bias exists in a participant when the participant chooses not to provide definitive responses⁹³ to 20% or more of the assessment questions.

Task 5. Assess model/human performance

(a) **Focus:** Inter-SME Consistency

(b) **Task:** Identify if there is agreement between SMEs on the level of performance for each subtask, task, scenario, and overall question.

(c) **Measure:** Inter-SME Consistency is achieved when the standard deviation of the normalized scores (0 to 1) for SME responses when observing and assessing the same behavior is less than or equal to 0.10 (10%).

Task 6. Assess model/human performance

(a) **Focus:** Intra-SME Consistency

(b) **Task:** Identify if each participant demonstrates the ability to maintain concurrence between the average of the participant's sublevel response scores and the level score.

(c) **Measure:** Intra-SME Consistency is achieved when the differences between a SME's mean sublevel assessment value and the SME's level assessment response is less than +/- 0.5, a difference in score equal to one of the scale nominal values.

Task 7. Assess model/human performance

(a) **Focus:** Intra-SME Consistency Impact

(b) **Task:** Identify if participants with intra-SME inconsistency demonstrate a level of inconsistency that changes the participant's sublevels to level results from Go to No-Go, No-Go to Go, Unknown to Go, etc.

⁹³ A definitive response is a "Go" or "No-Go" assessment of the subtask, task, scenario, or overall assessment question. "Not Applicable" or "No Opinion" responses are not definitive responses.

(c) **Measure:** Intra-SME Consistency Impact is achieved when the differences between a SME's mean sublevel assessment value and the SME's level assessment response does not change the overall assessment of the level from Go to No-Go, Go to Unknown, No-Go to Go, No-Go to Unknown, Unknown to Go, or Unknown to No-Go.

Task 8. Assess model/human performance

(a) **Focus:** Intra-SME Accuracy

(b) **Task:** Identify if each participant demonstrates the ability to maintain relative correctness with respect to a consistent scale dependent assessment key of each subtask, task, scenario, and overall assessment.

(c) **Measure:** Intra-SME Accuracy is achieved when the number of differences between a SME's assessment responses and the associated scale's key assessment values is less than 10% of the total number of assessed tasks when observing and assessing the same behavior.

Task 9. Assess model/human performance

(a) **Focus:** Intra-SME Accuracy Impact

(b) **Task:** Identify if participants with intra-SME inaccuracy demonstrate a level of inaccuracy that changes the participant's level assessment from Go to No-Go, No-Go to Go, Unknown to Go, etc.

(c) **Measure:** Intra-SME Accuracy Impact is achieved when the differences between a SME's assessment responses and the associated scale's key assessment values does not change the overall assessment of the level from Go to No-Go, Go to Unknown, No-Go to Go, No-Go to Unknown, Unknown to Go, or Unknown to No-Go.

APPENDIX E. EXPERIMENT MATERIALS

The following is a listing of equipment and materials required for the studies conducted at Fort Benning, GA.

1) Room:

- a. Chair(s) 22 for experiment #1; 29 for experiment #2
- b. Table(s) 13 for experiment #1; 29 for experiment #2
- c. PC (Dell Inspiron 8200, 2.0 MHz Laptop)
 - i. MANA (agent based model)
 - ii. Scenario(s)
 - 1. Practice
 - 2. McKenna Squad Raid #1
 - 3. McKenna Squad Defend Building
 - 4. McKenna Squad Raid #2
- d. PC (Dell Dimension 8200, 2.52 MHz Desktop)
 - i. MANA (agent based model)
 - ii. Scenario(s)
 - 1. Practice
 - 2. McKenna Squad Raid #1
 - 3. McKenna Squad Defend Building
 - 4. McKenna Squad Raid #2
- e. Computer Projector (Dell 3200MP)
- f. Printer (Epson , Color DeskJet)
- g. Screen – Display Surface (5' x 5')

- 2) Project Binder Containing:
 - a. Data Collection Sheet(s)
 - b. Terrain Sketch(es)
 - c. In-Brief Script
 - d. Study Script
 - e. Debrief Script
- 3) Data Recording:
 - a. Participant Books
 - i. Minimum Risk Consent Statement
 - ii. Participant Consent Form
 - iii. Privacy Statement
 - iv. Subject Personal Data Sheet
 - v. NEO Five Factor Inventory
 - vi. Model Evaluation Form(s)
 - vii. Terrain Sketch(es)
 - viii. Debrief Statement
 - ix. Scratch Paper
 - x. Model Interface Reference Poster
 - xi. Assessment Reference Poster
 - b. Camera (8mm)
 - c. Blue pens
- 4) Misc
 - a. Batteries (8mm Camera)
 - b. Batteries (AA) – projector controller

- c. Batteries (Laptop)
- d. Case of Printer Paper
- e. Clock (participant(s))
- f. Digital Camera
- g. Disposable Camera (back-up)
- h. Drop Box
- i. Envelopes (for debriefing handout and participant NEO FFI Summary)
- j. Fans (two for ventilation)
- k. Light bulbs (2 spares for 3200MP projector)
- l. Manila Folders (hold participant records)
- m. Map Markers
- n. Power Cord
- o. Power Strip
- p. Spare pens/markers
- q. Stapler
- r. Stop watch (researcher)
- s. Tape (Scotch & Packing)
- t. Three Hole Punch
- u. Visitor Briefing Book

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX F. STUDY ENVIRONMENTS AND SCENARIOS

The two primary studies for this research employ three environments and four scenarios. For training, model familiarization, and practice of assessment procedures, a fictitious environment and scenario are used. For data collection, the remaining two environments are based on the McKenna MOUT site developed in MANA, with three scenarios (one defensive and two offensive) generated for assessment. This appendix describes these environments and scenarios. Note that the top of all sketches and maps is north.

1. ENVIRONMENTS

The environments created for the MANA are built of color-coded bitmapped images, which the user employs to create and display background, elevation, and terrain images. Background images make presentation of entity animation easier to visualize by providing a context; they have no impact on behaviors [are they grayscale or color? Is F1 an example? I didn't understand why they would have no impact if they show buildings, etc, because obviously the presence of a door or window would influence decisions]. Elevation is represented by a grayscale image—the higher the elevation, the lighter the value. Terrain complexities are shown via five-color bitmaps. This study employs only background and terrain images; due to the limited accuracy of the model's elevation maps and the relatively uniform terrain at the McKenna MOUT site, elevation is not shown.

For the underlying terrain in MANA, each pixel is one terrain square. Each terrain square can be occupied by one or more agents, except if its color indicates terrain that entities cannot penetrate (e.g., walls and cliffs). There are five terrain representations, distinguished by color: unrestricted terrain; high-speed avenues of approach; restricted terrain; severely restricted terrain; and impassable terrain.

In MANA, unrestricted terrain is normally shown as black, brown, or tan, but it can be any color except yellow, light green, dark green, or grey. Also known as “plain” terrain, unrestricted terrain provides no special opportunities for movement, cover, or concealment.

High-speed avenues of approach, or “easy-going” terrain, are colored yellow. Though these avenues provide neither cover nor concealment, they are attractive to agents for offering ease of movement, and are commonly roads or trails.

Restricted terrain reduces an entity’s speed but provides cover and concealment. Shown as light green, it consists of nominal bushes and scrub. For urban scenarios, light green is also used to indicate the reduced visibility and circumscribed movement of entities inside buildings.

Severely restricted terrain, which would include heavy vegetation, is represented by dark green. This terrain greatly reduces rate of movement and provides deep cover and concealment. In this study, the dark green of severely restricted terrain is used both for the wooded areas around buildings and for window openings.

Finally, impassible terrain, such as walls, are represented by gray. No entity can occupy gray areas, and the line of sight between entities on opposite sides is assumed to be blocked. [GALL 03]

Below are the practice, offensive, and defensive environments created for the studies. Each represents an urban setting, with differing levels of fidelity. The resolution of each environment is 1000 by 1000 pixels.

a. Practice Environment

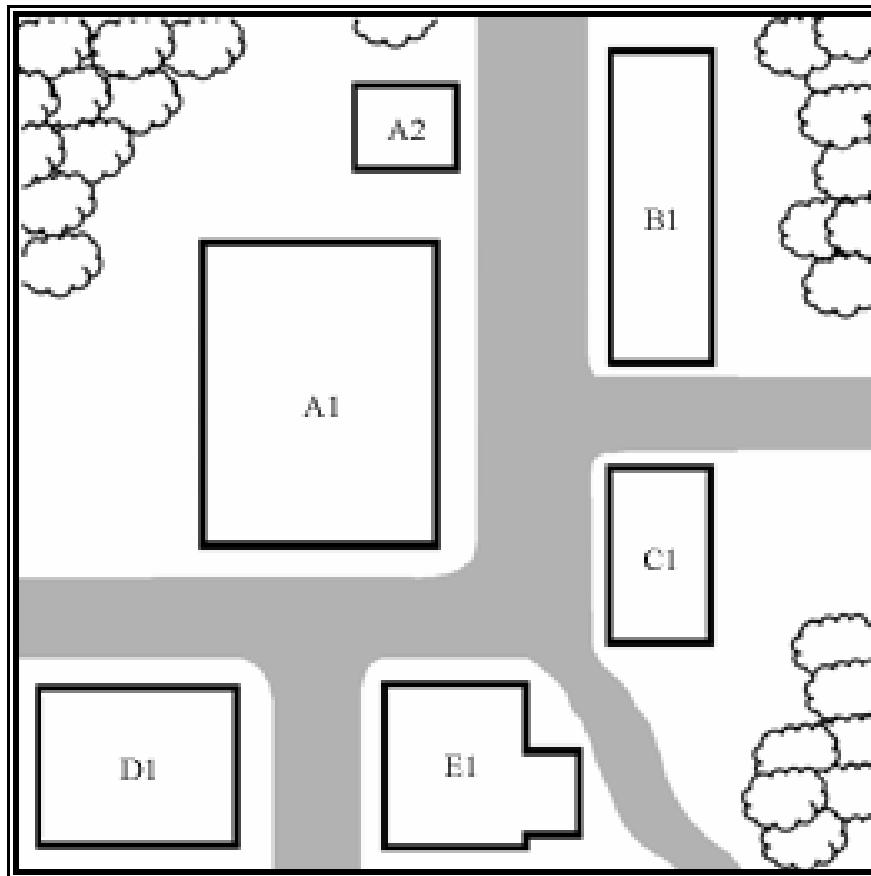


Figure F1. Training and Practice Environment Building Numbers.

The practice environment consists of six buildings situated around a meeting of five roads and trails. The buildings are numbered by block, sequentially from west to east and north to south (Figure F1).

Figure F2 depicts the terrain underlying the crossroads sketch (Figure F3) displayed for the participants. The buildings confront severely restricted terrain to the northwest, north, northeast and southeast. Two of them, buildings *AI* and *DI*, have superstructures (Figure F1); only the first floor layout, however, is used for the scenario.

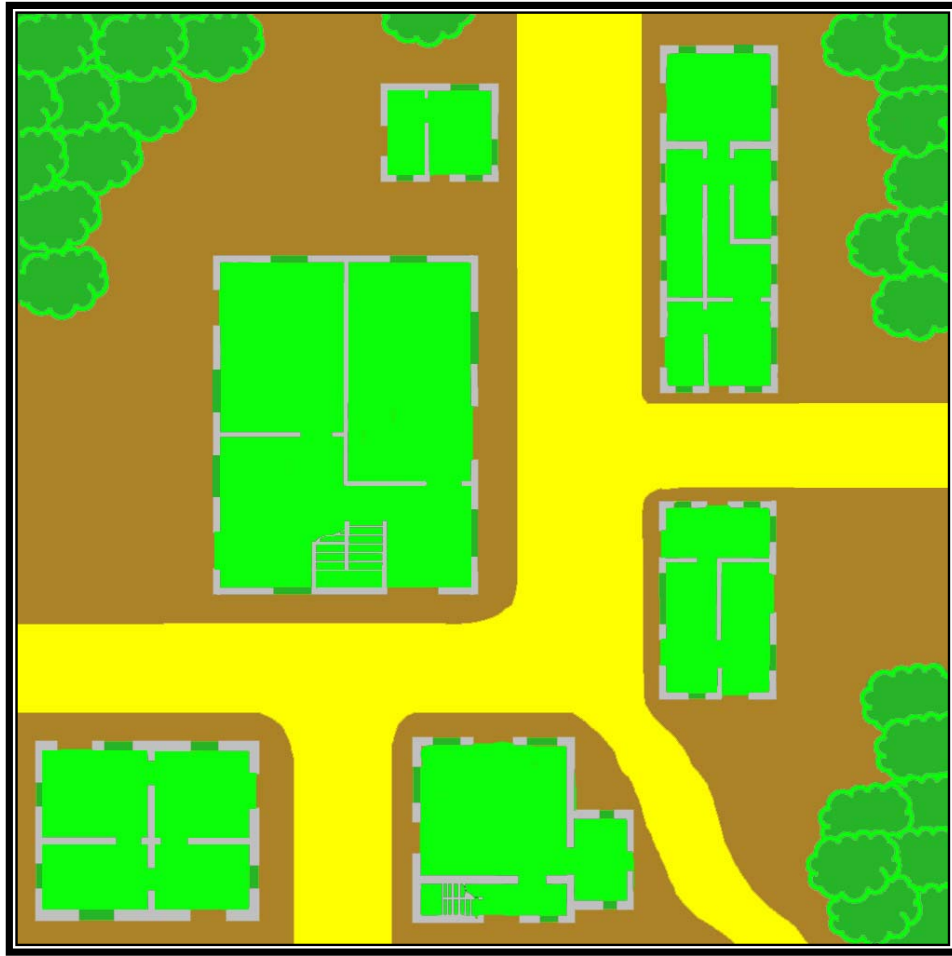


Figure F2. Training and Practice Environment Terrain Sketch.

The crossroads floor-plan sketch (Figure F3) provides the background. The shades of gray make it easy for participants to identify forces and agent interactions. The scheme also provides visual cues concerning the area of operations.

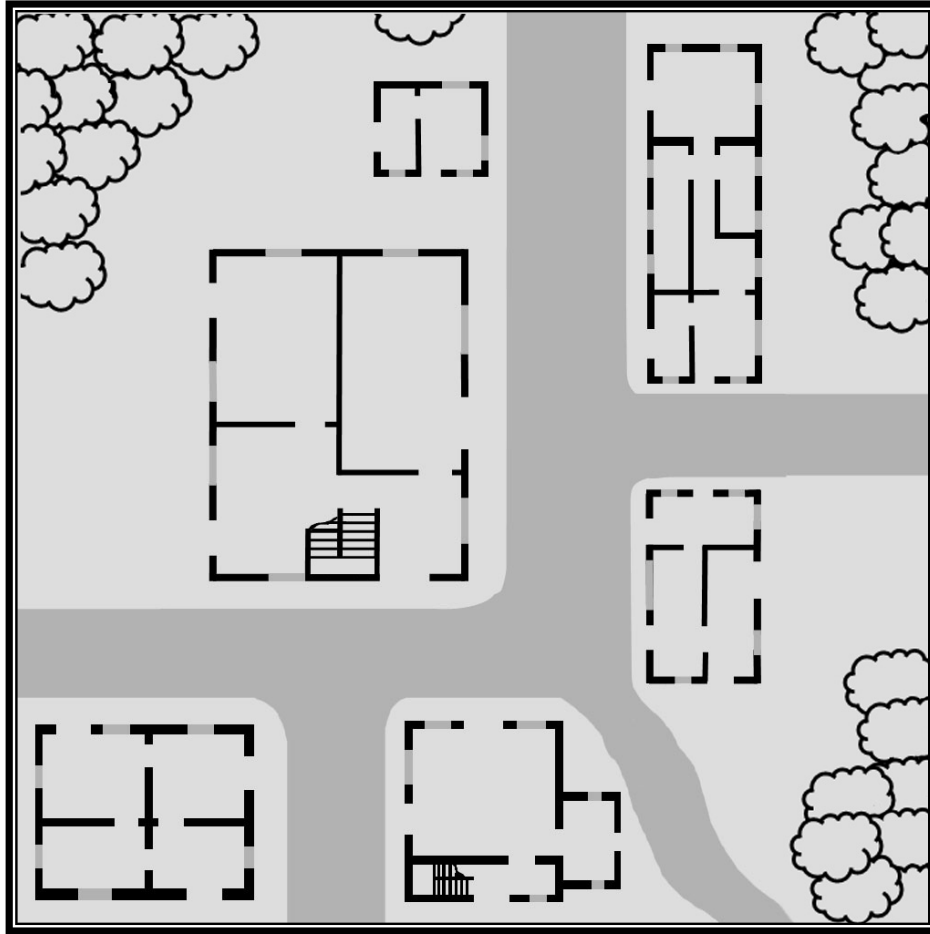


Figure F3. Training and Practice Environment Floor Plan (Display Sketch).

Walls and stairways are displayed in black, windows and streets in dark grey. The light grey of interiors, doorways, and open terrain allows increased visibility of agent animations. Mines, wire obstacles, and tree lines (Figure F13) are black.

b. Offensive Test Environment

The offensive test environment for these studies is the McKenna military operations in urban terrain (MOUT) site, Fort Benning, GA (Figure F4). This environment consists of twenty-eight buildings, ranging from a storage shed to a three-story hospital. The buildings are arrayed so that no road offers an unobstructed line of sight from one side of the village to the other. The environment has a church in the center of town (Figure F4, building 33), a bank (building 32), a hospital (building 41), a police station (building 42), single and multiple-family dwellings, and a small garrison compound (Figure F5, buildings *J1* and *J2*).

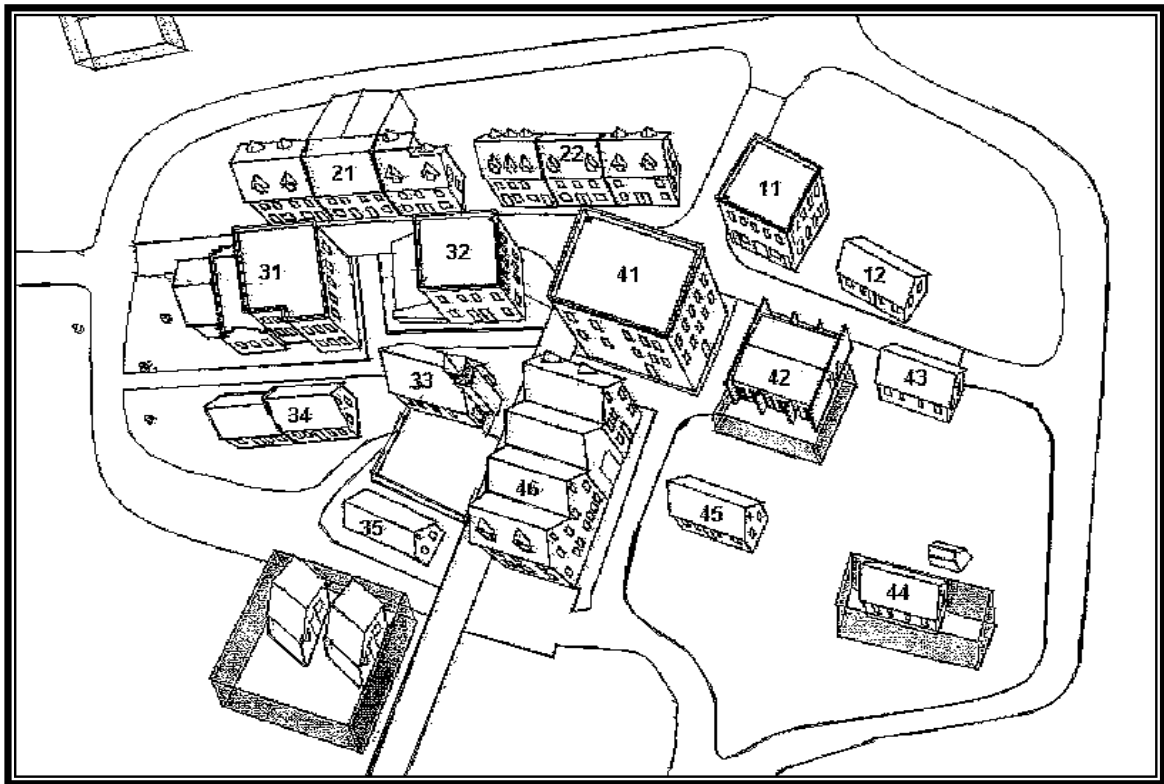


Figure F4. Offensive Test Environment Sketch From [STAT 03]

The buildings are situated on nine city blocks and referenced based on their block and position within the block. Blocks are ordered alphabetically, starting from the village's northwestern corner and ending at the southeast (Figure F5). Within each block, buildings are designated numerically, again starting from the northwestern corner of each block and progressing to the southeast. Building **D1** is the village bank, **E1** is the hospital, **F1** is the police station, and **G3** is the church.

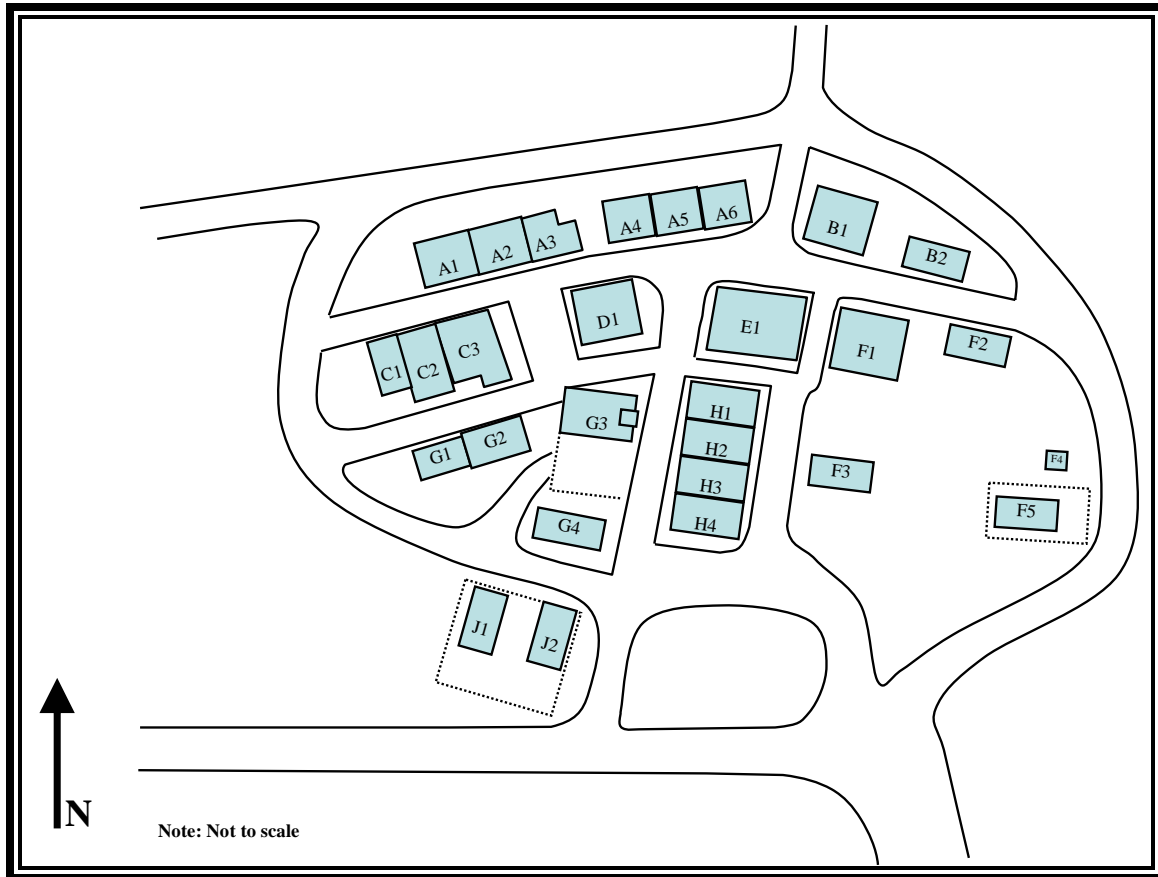


Figure F5. Offensive Test Environment Building Numbers

The underlying terrain for the McKenna MOUT site uses the same five-color scheme described for the terrain in the practice environment. Figure F6 illustrates the terrain image used for the McKenna offensive scenarios. Severely restricted terrain borders the trail encircling the village. While fifteen of the buildings have superstructures—*A1-A6*, *B1*, *C3*, *D1*, *E1*, *G3*, and *H1-H4* have two stories and *E1* has three stories (Figure F5)—only the first-floor layouts are used. Most buildings are depicted at lower fidelity than those of the practice terrain. Few doors, windows, or interior walls are depicted, because of the greater scale of the environment.



Figure F6. Offensive Test Environment Terrain Sketch.

To provide additional avenues of approach and protection for agents operating in the area of central interest, interior walls for the first floor of buildings on the *H* block are included, but their placement is not as detailed as those in the defensive environment.

They serve to increase similarity between the offensive and defensive environments while maintaining a level of distinction between the two. As with the practice environment, the terrain resolution for the offensive environment is 1000 by 1000 pixels.

c. Defensive Test Environment

The defensive test environment for these studies is the McKenna MOUT site, Fort Benning, GA, consisting of the four buildings in the **H** block of McKenna (Figure F5), the streets and trails surrounding the buildings, the front of the two buildings immediately west of the complex, and the building directly east.

Figure F7 illustrates the floor plans involved. This image is the display background for the defensive scenario; only first floors are used. Building **H2**, center and second from the top, is a garage with an opening facing east. The church (Figure F5, building **G3**) is northwest of the Figure F7 floor plan.

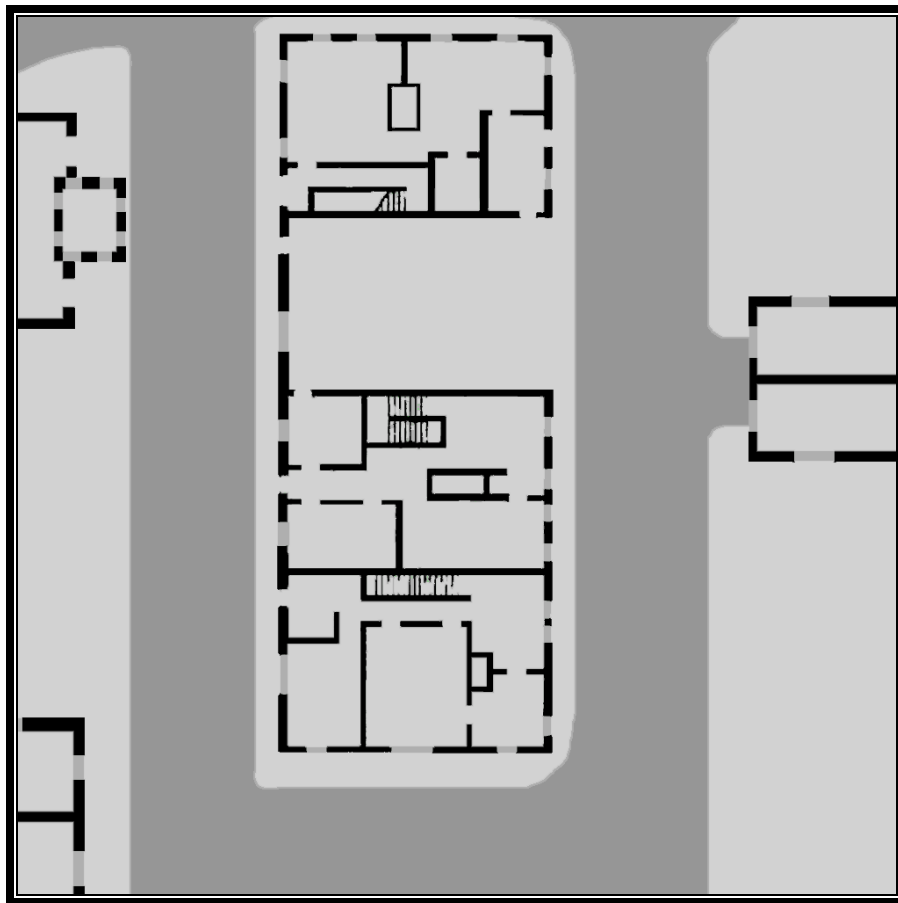


Figure F7. Defensive Test Environment Floor Plan (Display Sketch).

The primary terrain image for the defensive environment uses the same five-color scheme describe for the terrain in the practice and offensive environments. Figure F8 shows the terrain characteristic residing beneath the floor-plan sketch. As with the practice and offensive environments, the terrain resolution for the defensive environment is 1000 by 1000 pixels.

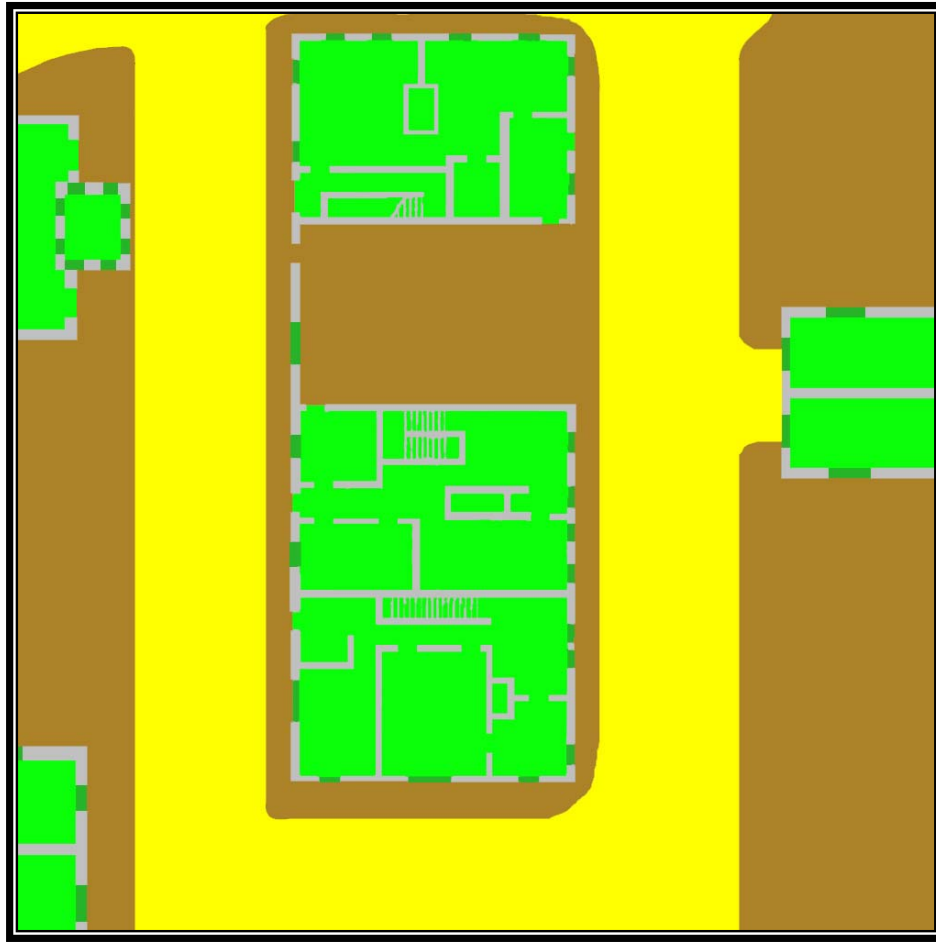


Figure F8. Defensive Test Environment Terrain Sketch

This depiction of the south-central section of the MOUT site displays higher fidelity than the terrain in the offensive environment, equal to that of the practice scenario (showing all windows, doorways, stairwells, and walls of first floors).

2. SCENARIOS

The participants undergo four scenarios: one practice and three test. The practice scenario is executed on the training-and-test-scenario environment. The scenarios used

for data collection are executed on the McKenna MOUT site environments. Each scenario is executed by MANA Version 3.0.13. The resulting animations are displayed to participants using the MANA interface and projected on a 5' x 5' white screen or whiteboard.

Due to visualization limitations of MANA, the need for simplistic scenarios, and the focus of the Oct-Nov 2002 Natick study, strictures were placed on the use of indirect-fire weapons, artillery, mortars, and grenade launchers. Participants learn that the blue forces will follow rules of engagement (ROE) to limit collateral damage and preserve life where possible. The following is the ROE for all blue-force elements:

- Take all steps necessary and appropriate for your unit's protection.
- Use the minimum necessary force to control the situation. Place rifles in single-shot mode to reduce fratricide, civilian casualties, and excessive use of ammunition.
- To reduce friendly casualties and damage to buildings, use concussion grenades within the boundaries. If you deploy a dummy flash bang, there will be a two-second delay before detonation. Concussion grenades will incapacitate personnel for approximately five seconds—swift movement to secure enemy personnel is essential.
- Follow standard MOUT tactics, techniques, and procedures (TTPs).
- Take measures to minimize risk to civilians without endangering the unit.
- Return fire directly to its source, not spraying a general area (use single-shoot selection for rifles).
- Cease firing when the threat is over.
- Anyone trying to surrender is allowed to do so.
- Treat civilians and property with respect.
- Use white phosphorous (WP) as needed in the vicinity of the town to aid in isolating objectives. The battalion commander must authorize requests for the use of indirect fire within the town.
- Do not use artillery inside the town.

The four scenarios are described below. First is the practice scenario, followed by the two offensive scenarios and the defensive.

Practice Scenario

Before viewing the practice scenario, participants draw the defensive positions of the squad members using standard symbology from FM 101-5-1 (Figure F9). Participants have unlimited mines and wire obstacles (Figure F10), but are limited to a given number of personnel and weapon systems for the defense of a building's first floor: namely, to six M16A2 assault rifles or M24 sniper rifles, two M203 grenade launchers affixed to M16A2 assault rifles, two M249 squad automatic weapons (SAWs), and four M136s, AT-4 84mm rocket launchers (Figure F11). Figure F12 illustrates how one might place these weapons systems in building *C1* to provide 360-degree protection.

Mines	
Antipersonnel (AP)	
Antitank (AT)	

Wire Obstacles	
Unspecified	XXXXXXXXXX
Triple Stand Concertina	

Weapons		
M16A2/M24 (6)		
M203 (2)		
M249 (2)		
AT-4 (4 located w/M16)		
60mm Mortar (0)		

Figure F9. Training and Practice Environment Sketch Symbology From [DEPA 97a]

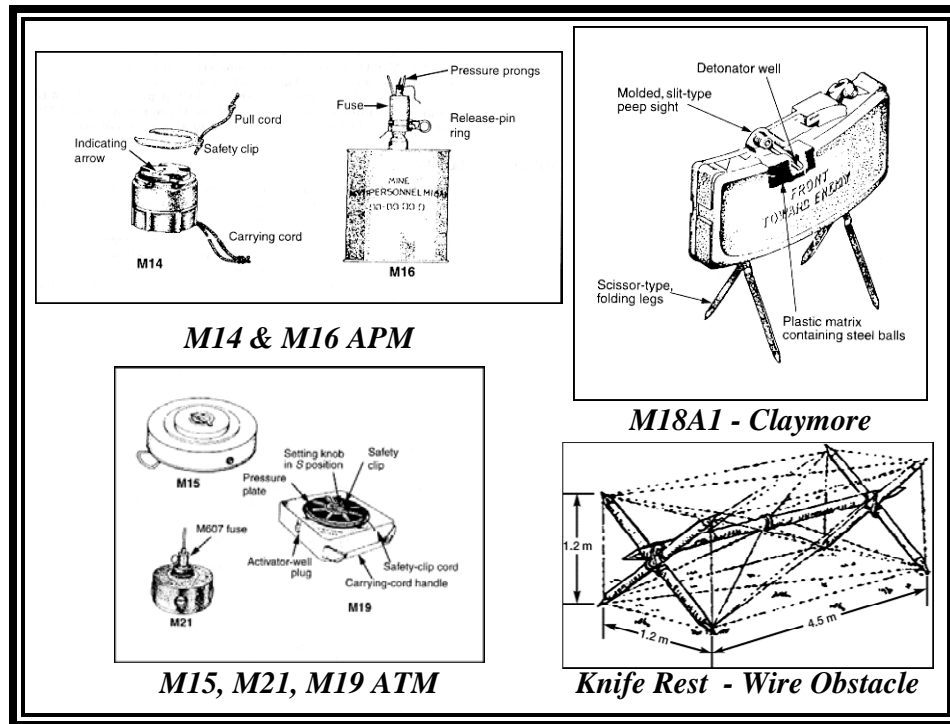


Figure F10. Mines and Wire Obstacles From [US L 99] [SOLD 03]

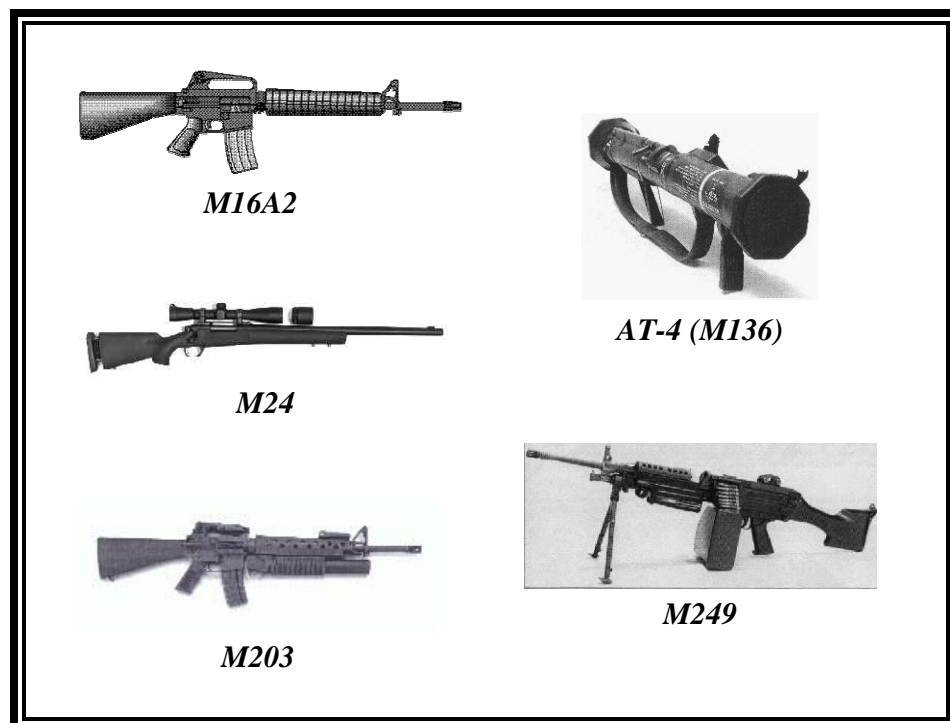


Figure F11. Weapons Systems From [US L 99]

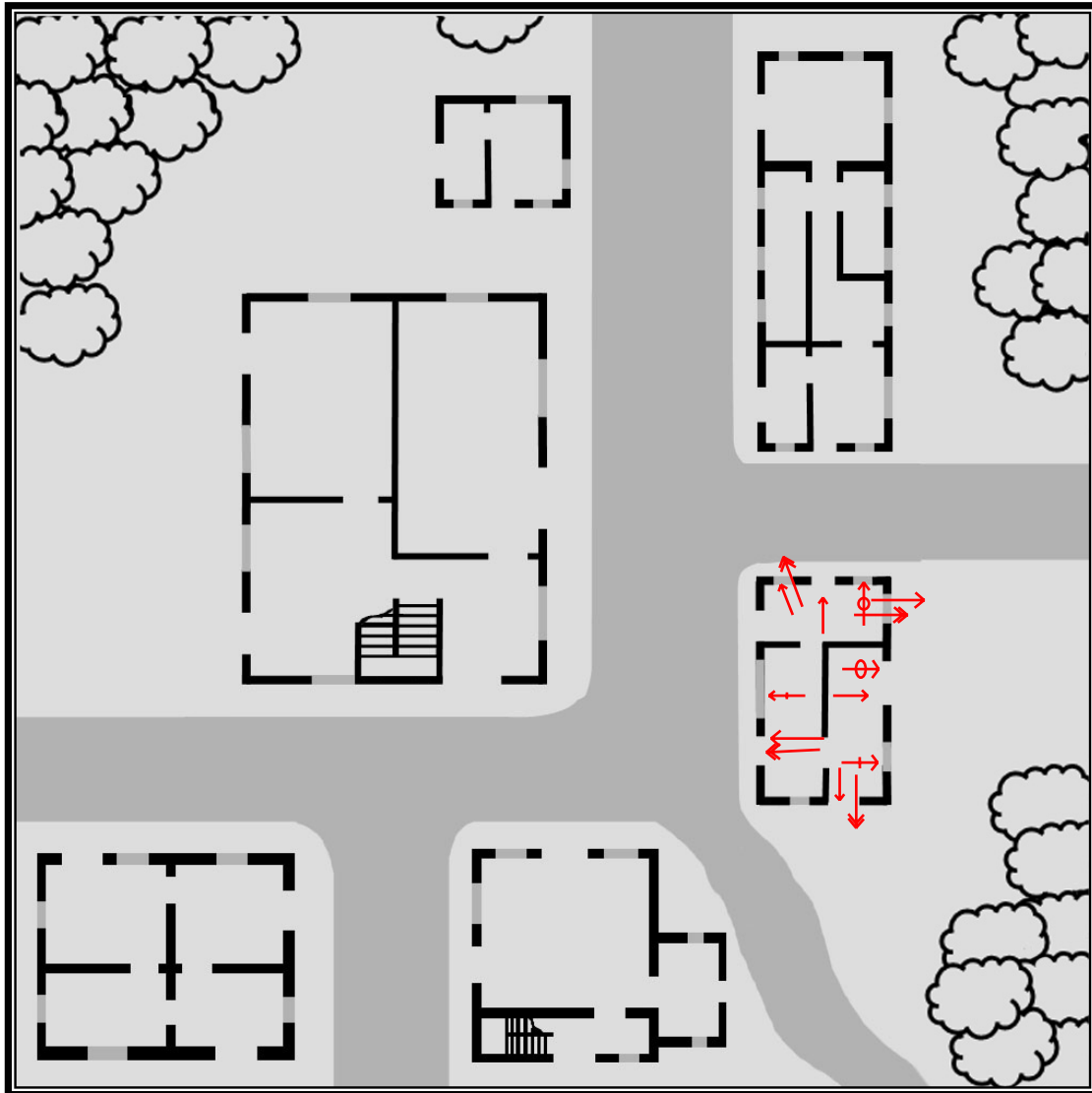


Figure F12. Training and Practice Environment Floor Plans with Example Defensive Positions for Building *E1*

In the practice scenario, participants assess the performance of a squad defending building *A1* (Figure F1). The squad, consisting of ten personnel, is defending against an enemy squad of nine, who attempt to reach building *A1* from the north of the MANA display (Figure F13). The squad leader is positioned in the center of the northwestern room with one of his teams. The second team is defending the southwestern and northeastern rooms. There is also a man in the southeastern room. The defenders have placed antipersonnel mines and wire obstacles outside all windows of building *A1* and

blocked the main eastern entrance with a wire obstacle. Civilian personnel are found in all buildings except *A1* and *A2*. Figure F13 shows the initial state of forces in the practice scenario.

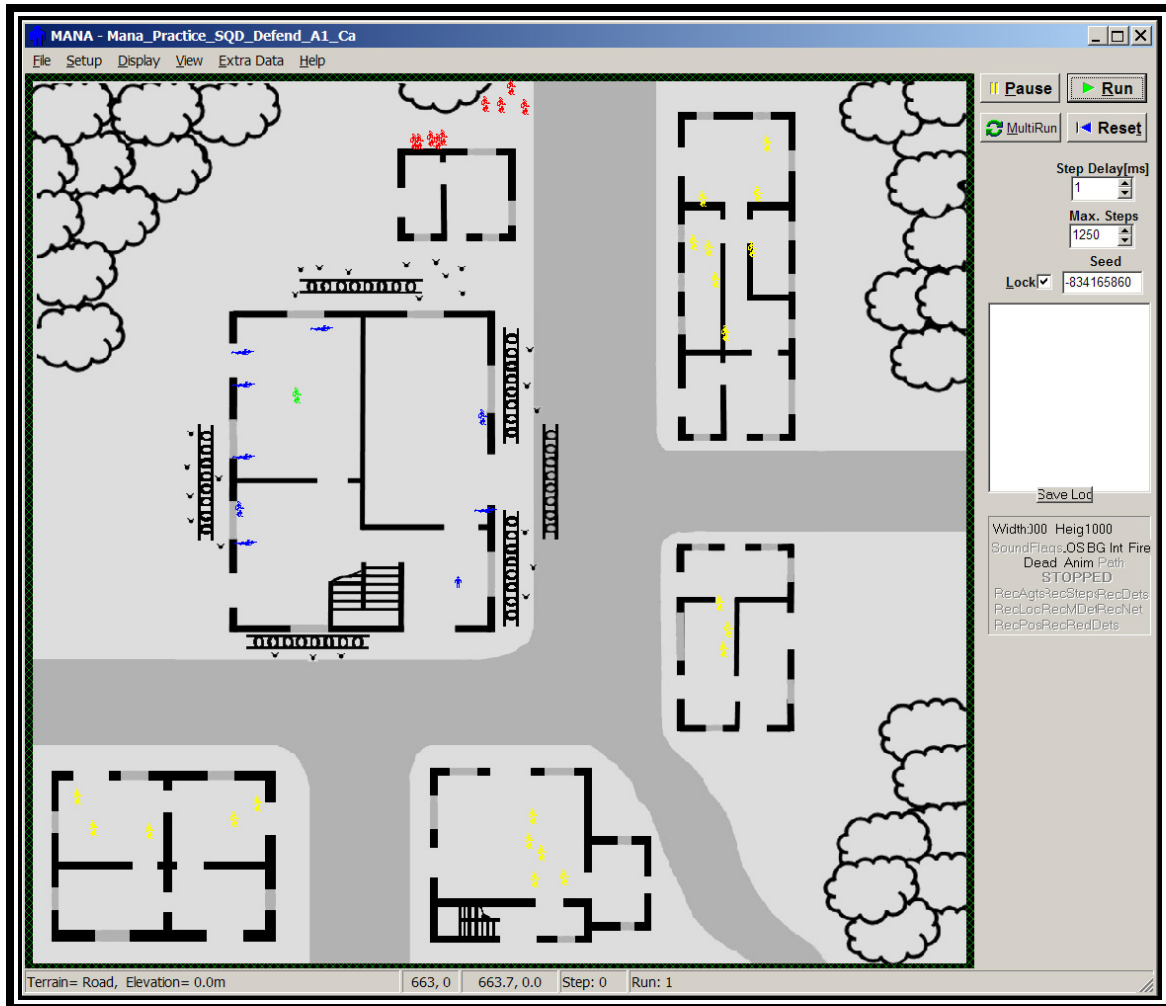


Figure F13. MANA Display of Training and Practice Environment; Initial State for Defensive of Building *A1*

During the scenario run, civilians in the buildings at the east and south of the area of operations seek cover and concealment indoors (Figure F14). The red force approaches *A1* from the north of building *A2* (Figure F13). As its lead team rounds the northwestern corner of *A2*, they are seen and come under fire by the blue forces guarding the

northwestern window of *A1*. The blue-squad leader moves one of the entities guarding the northwestern door (initially looking west) to the northwestern window to repel red forces.

The second blue-team leader, positioned in the southwestern room of building *A1*, moves to the northeastern room, where he finds that the entity guarding the door has maintained his position, but the entity guarding the window facing the street eastward has moved to the northeastern window to counter red forces coming through building *A2*.

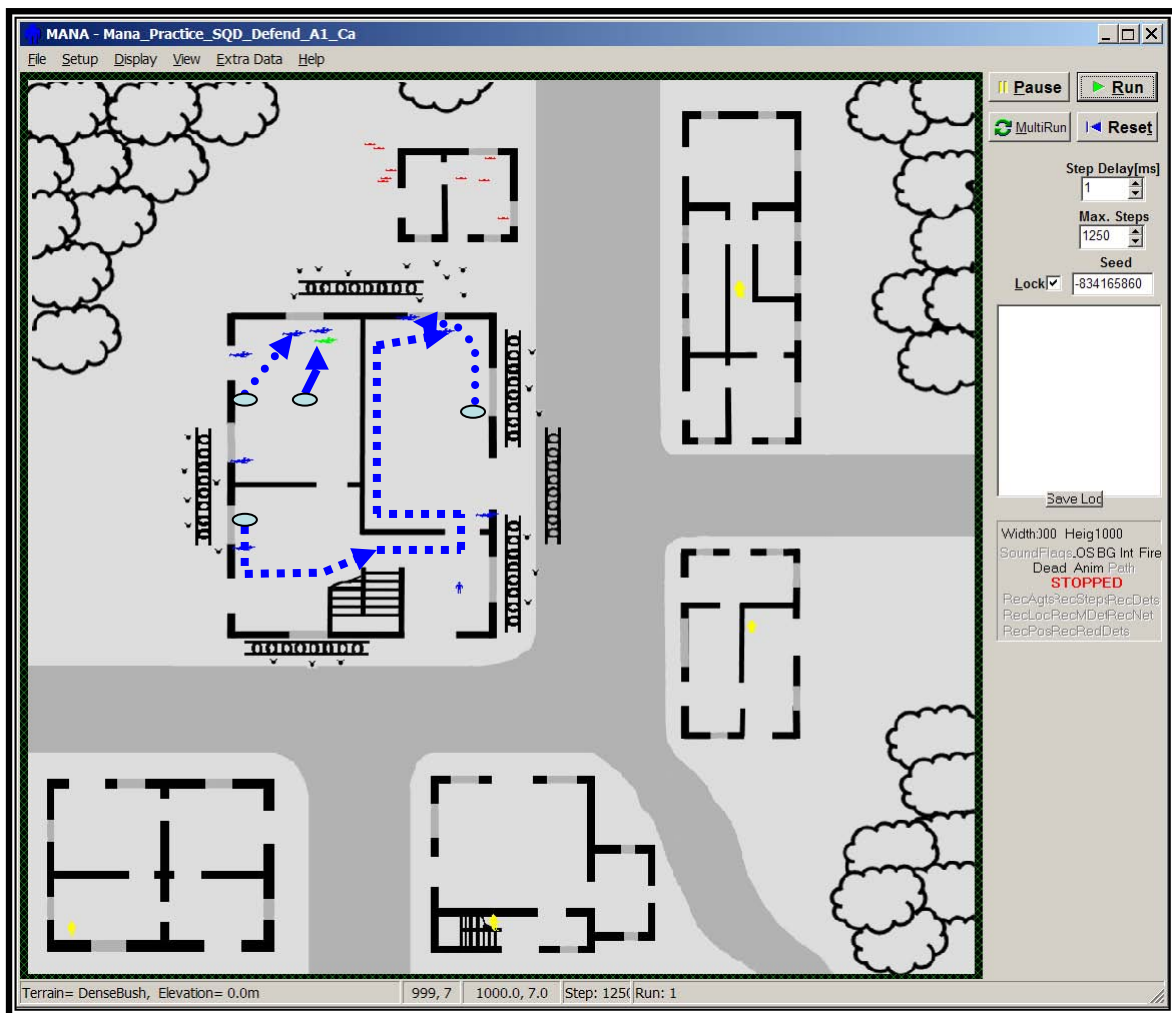


Figure F14. MANA Display of Training and Practice Environment; Final State and Blue-Force Routes for Defensive of Building *A1*

For the practice run, operating at near real-time speeds and without interruption, the scenario lasts about twenty-two seconds, and thirteen seconds when played faster than real time. The blue force destroys all red entities without squad or civilian losses (Figure F14). The extensible mark-up language (XML) file storing the scenario is 2.191MB in size.

a. Offensive Scenario #1

The offensive scenario is based on a study performed by Statkus, Sampson, and Woods in the fall of 2002 at the McKenna MOUT Site. The Natick study examined squad-sized units conducting operations in an urban environment in an attempt to measure the effects of situational awareness on troop movement and decision making. Research personnel collected “sample data on movement patterns and cognitive thinking processes that effect movement behaviors in MOUT combat scenarios.” [STAT 03] The environment, weapons systems, rules of engagement, and some of the forces and force positions are used as bases for the offensive scenarios.

The enemy is positioned inside the village of McKenna (Figure F15). It consists of three squad-sized elements with a military intelligence (MI) section. The MI section is located in building **H4** along with a POW whom the friendly forces seek to rescue. One enemy squad screens to the north of the village. A second squad is garrisoned in the compound to the southwest of the village, buildings **J1** and **J2**. This second squad is the quick-reaction force (QRF), designated to reinforce defenses threatened by blue forces or to counterattack blue elements. The third squad patrols the village in team-sized elements, two to three personnel per team. The enemy situation established by Natick is used for the third squad and MI section in both MANA offensive scenarios.

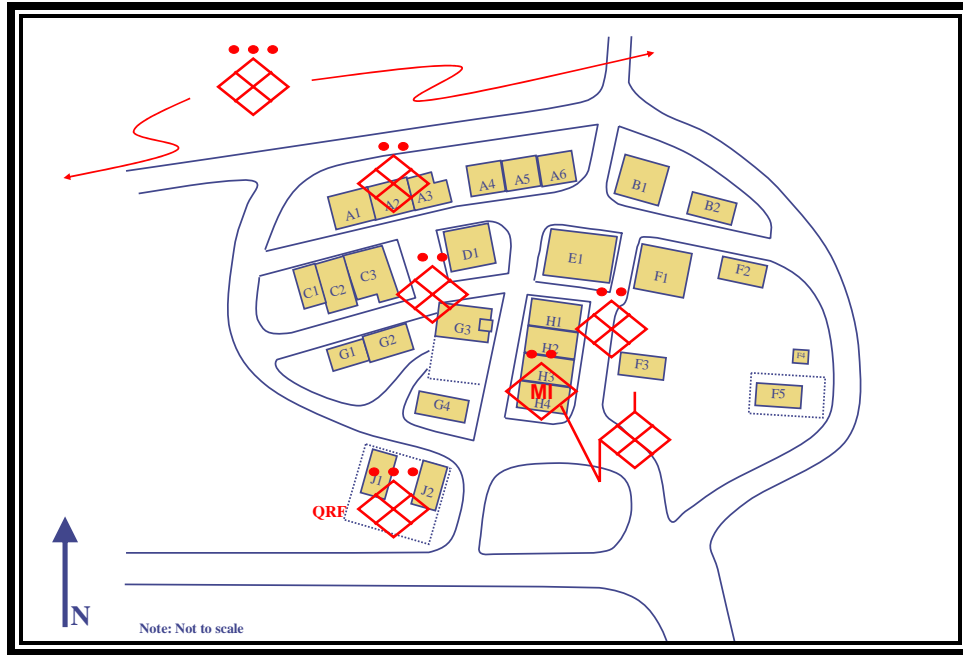


Figure F15. McKenna Offensive Scenario Templated Enemy After [STAT 03]

The friendly forces consist of a battalion-sized element moving into the village and its surroundings to allow a squad element to sweep in and rescue the POW (Figure F16). One company of infantry secures Objective (OBJ) Fuse, the northern border of the village, thus neutralizing the enemy squad to the north. A second infantry company fixes the enemy QRF to the south, OBJ Dynamite. The third company approaches the village along routes Dime and Quarter to secure a foothold to the northwest of the village (OBJ Dime and OBJ Quarter), allowing a squad element to infiltrate the village and rescue the POW (OBJ Vault).

To maintain simplicity for the studies and limit the participants' required viewing time, the scenario portion modeled and displayed for assessment consisted of the enemy squad defending the area around building **H4** and the blue squad infiltrating the village to secure and extract the POW (OBJ Vault).

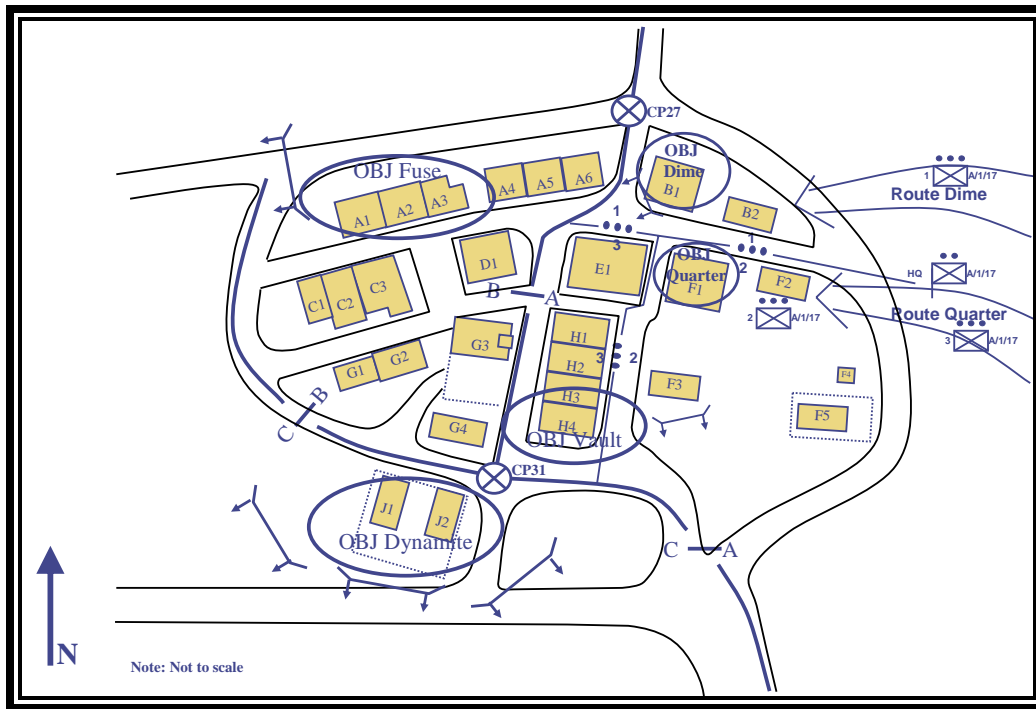


Figure F16. McKenna Offensive Scenario Blue-Force Graphics

Prior to seeing the MANA display of the first offensive scenario, participants are asked to sketch the route they would follow through the village if they were the squad leader assigned the task of rescuing the POW, using standard symbology from FM 101-5-1 for maneuver operations (Figure F17).

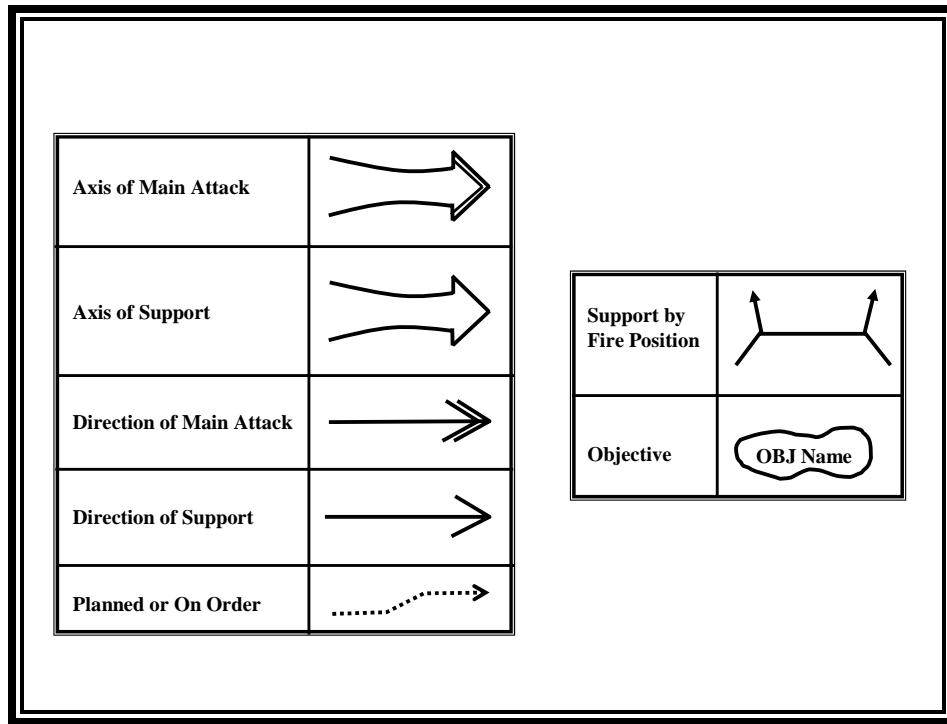


Figure F17. Offensive Sketch Symbolology From [DEPA 97a]

The blue forces operate on data gathered from the real-world performance of the first squad tested during the fall 2002 Natick study, *Human Science/Modeling and Analysis Data Project: Situation Awareness Effects on Troop Movement and Decision Making Data Collection Effort*. This squad approaches from the north along building **A6** (Figure F18). The squad is in a column formation, with the squad leader behind the lead team. They cross the intersection to the alley between **E1** and **F1**. Moving down the alley, they cross another alley to reach the eastern side of the **H** complex, killing an enemy stationed in the western alley between the **H** complex and building **E1**. The squad continues along the eastern side of the **H** complex to reach the southern edge of building **H4**, unaware it has lost two men to a second enemy gunman, the secondary sniper. Progressing along the southern edge of building **H4**, the squad encounters enemy personnel near the church before fighting their way into **H4's** western entrance.

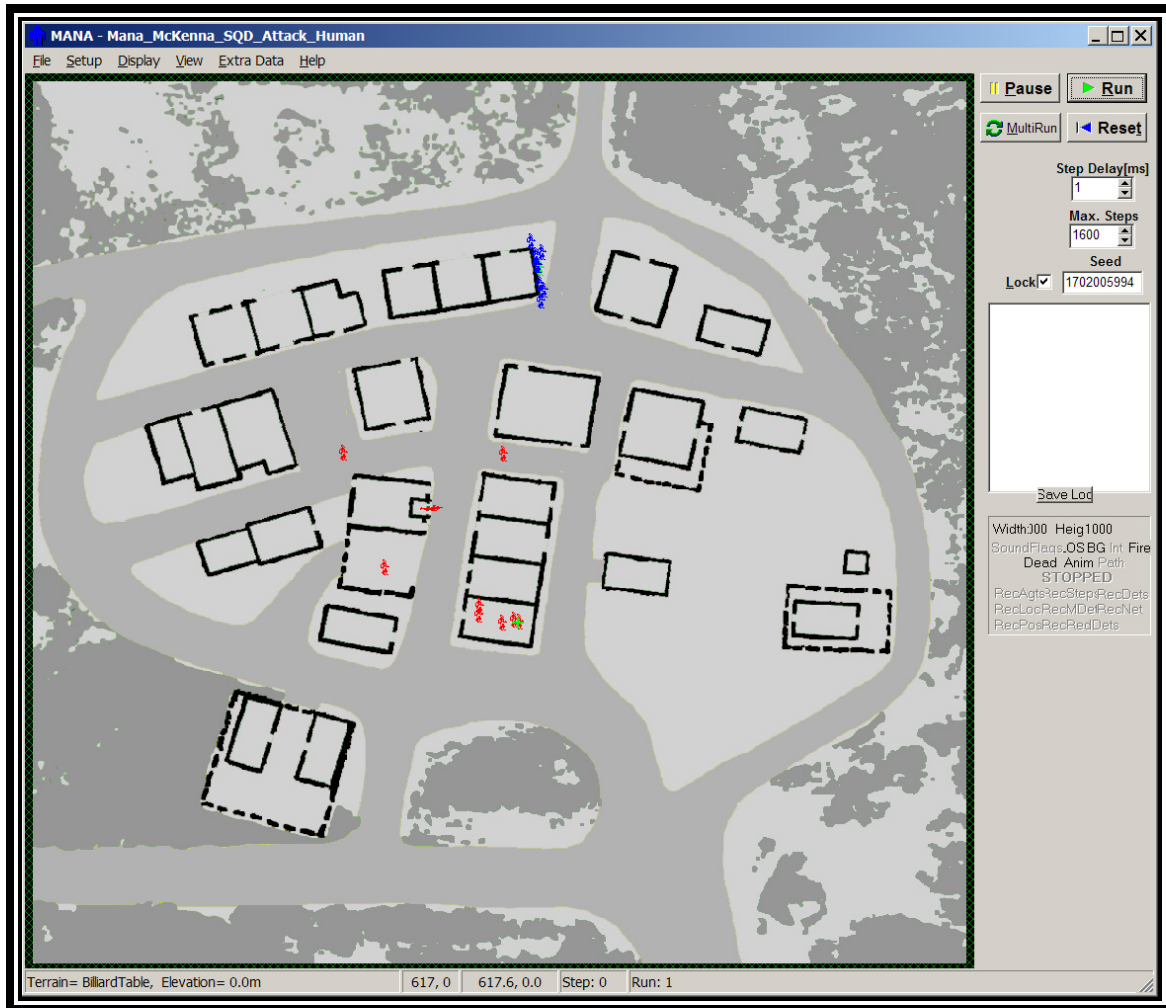


Figure F18. MANA Display of McKenna Offensive Scenario #1; Initial State

Figure F19 shows the route the real-world squad took through the environment for Movement Scenario 1 and where they encountered the enemy and lost personnel. Figure F20 shows the route taken by blue-force members in the MANA replay for Scenario #1, engagement points, and the end state of the scenario.

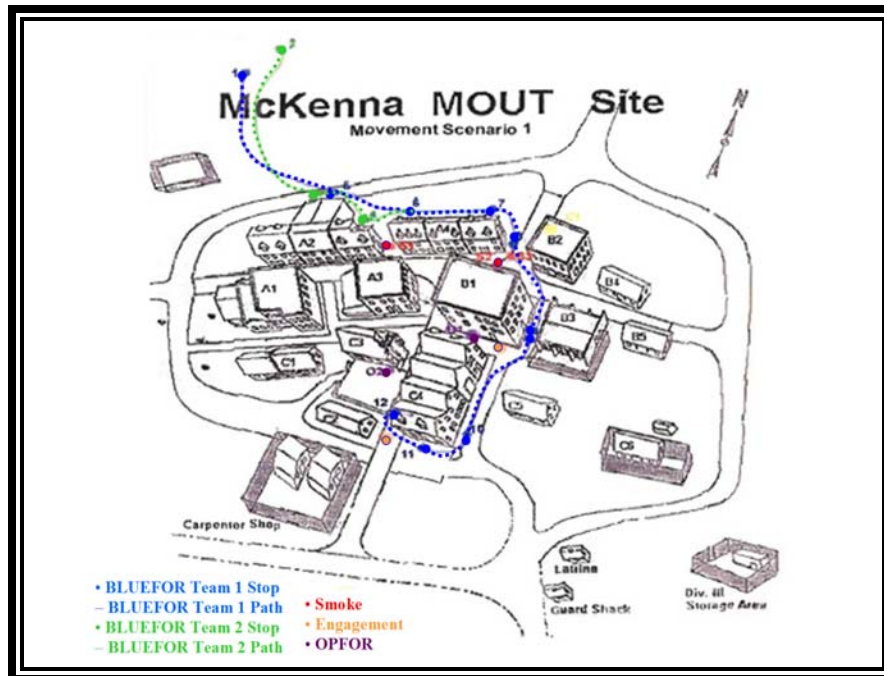


Figure F19. Natick Study; Movement Scenario 1 From [STAT 03]

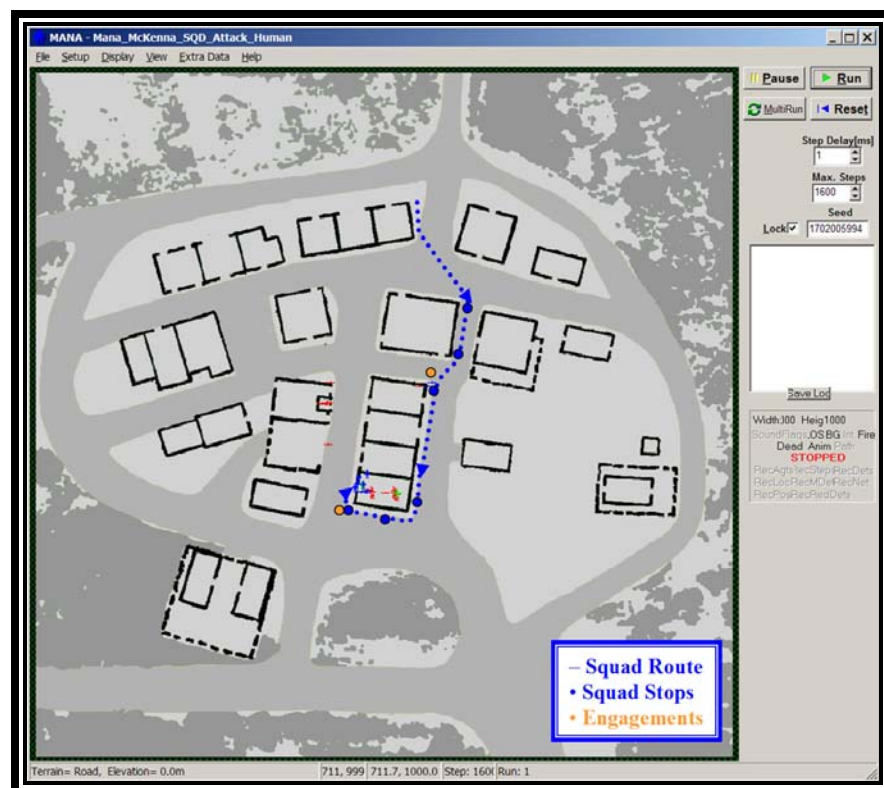


Figure F20. MANA Display of McKenna Offensive Scenario #1; End State, Blue Force Routes, and Engagement Locations

This scenario is the first to be assessed by participants during the data-collection phase of the study. Half the participants are informed the scenario is based on real data from a McKenna MOUT experiment; the other half are told the scenario is computer generated.

When run at near-real-time speed, the first offensive scenario lasts nearly nineteen seconds; at faster-than-real-time speeds, it runs seven seconds. The red forces lose four personnel and the blue squad lose two. The XML file storing this scenario is 1.076MB in size.

b. Offensive Scenario #2

The second offensive scenario uses the same enemy situation and ROE as the first, but differs in the route the infiltration squad takes into the village. The squad approaches from the east in a column of two team wedges with the squad leader positioned behind the lead team (Figure F21).

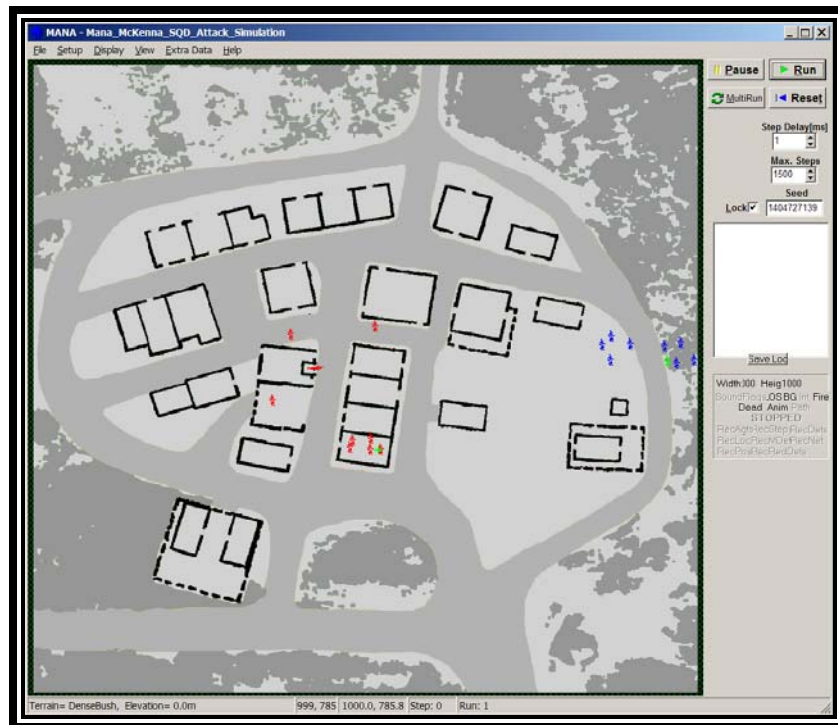


Figure F21. MANA Display of McKenna Offensive Scenario #2; Initial State

Once the squad reaches the first building on the eastern edge of the village, it changes formation to teams in column and continues moving west toward the alley between the **H** building complex and building **E1**. During this maneuver, the lead team encounters and destroys a red-force sentry at the northeastern corner of building **H1**. At the northwestern corner of **H1**, the squad encounters the sniper in **G3**. The lead team lays down a base of fire while the trail team maneuvers through building **H1** to gain position on the sniper (Figure F22). The squad kills the sniper and continues along the western wall of the **H** complex before it enters building **H4** through the western entrance.

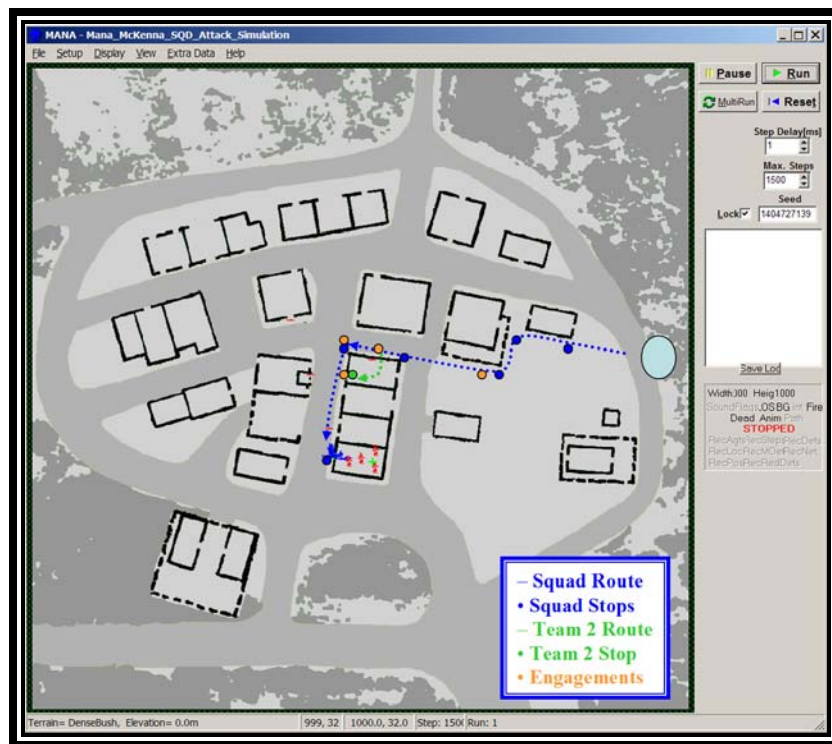


Figure F22. MANA Display of McKenna Offensive Scenario #2; End State, Blue Routes, and Engagement Locations

At near-real-time speed, this scenario lasts nearly 20 seconds. When run at faster-than-real-time speeds, this offensive scenario runs at eight seconds. The red forces lose four personnel and the blue squad lose none. The XML file for this scenario is 1.02MB.

This second offensive scenario is the last scenario to be assessed by participants during the data-collection phase of the study. As with the first offensive scenario, half the

participants are told the scenario is based on data collected at the McKenna MOUT site in 2002 and half are told it is computer generated.

c. Defensive Scenario

The defensive scenario has the same background as the two offensive scenarios, but the participants now assess the performance of the blue-force squad defending building **H4**. All members of the defensive force are in building **H4** and civilians are in buildings **H3** and **H1**. The red forces approach from the north along the building's eastern edge. Figure F23 shows the position of entities prior to the start of the scenario run.

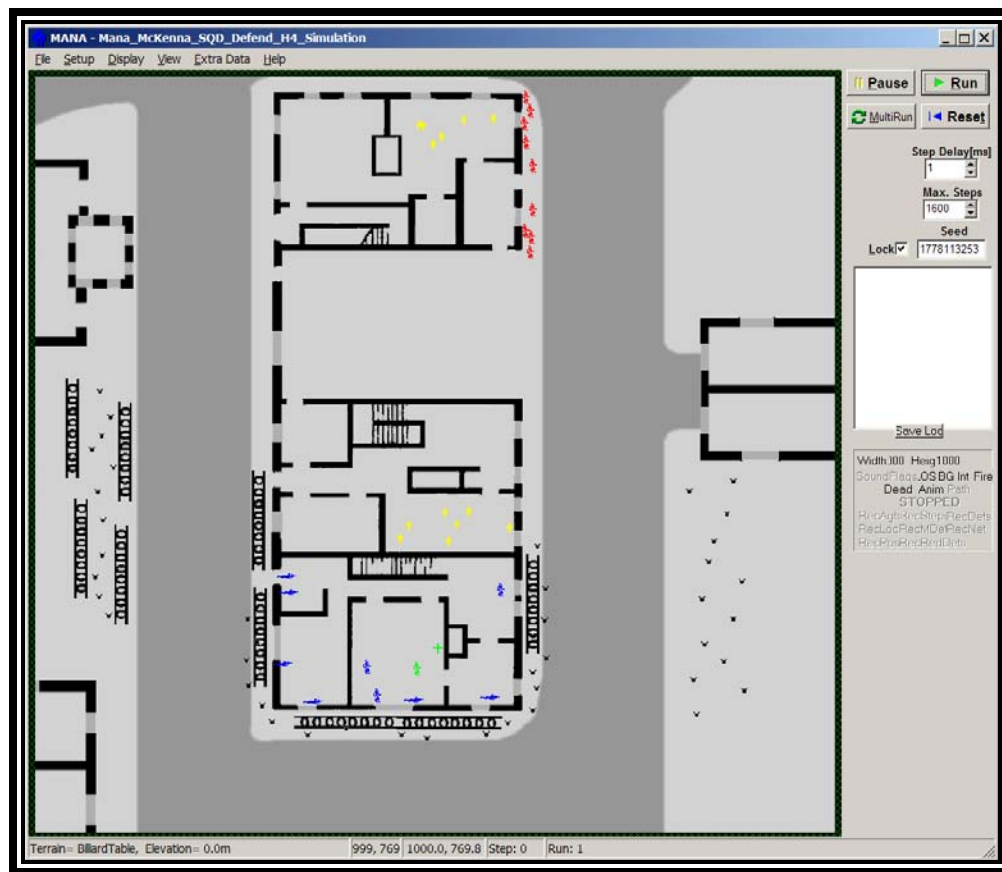


Figure F23. MANA Display of McKenna Defensive Scenario; Initial State

The squad is responsible for the defense of the first floor, with its primary mission keeping a piece of captured enemy equipment from the red force. As in the practice

scenario, participants are asked to sketch how they would mount a defense, using the same resources available in the practice scenario (Appendix E.2.a., Practice Scenario).

The blue-squad leader is positioned in the center of building **H4** with one of his teams and the captured equipment, designated by a “+.” The team, located partly in this room, is also defending the southeastern room. The second team is defending the main entrance to the building (the northwestern room) and the southwestern corner of the structure. A lone blue-force member is in the northeastern room by the stairs. The squad has emplaced antipersonnel mines and wire obstacles in front of all the windows of building **H4**. It has also placed an antipersonnel minefield in the eastern garden and a mixed wire-and-antipersonnel minefield obstacle in the church parking lot to the west.

The red forces approach along the eastern wall, moving south (Figure F24). The civilians in **H3** and **H1** begin to move to the center of the buildings once they identify red forces moving along the exterior walls. The blue-force member in the northeastern room of building **H4** detects the red force approaching from the north and notifies the squad leader.

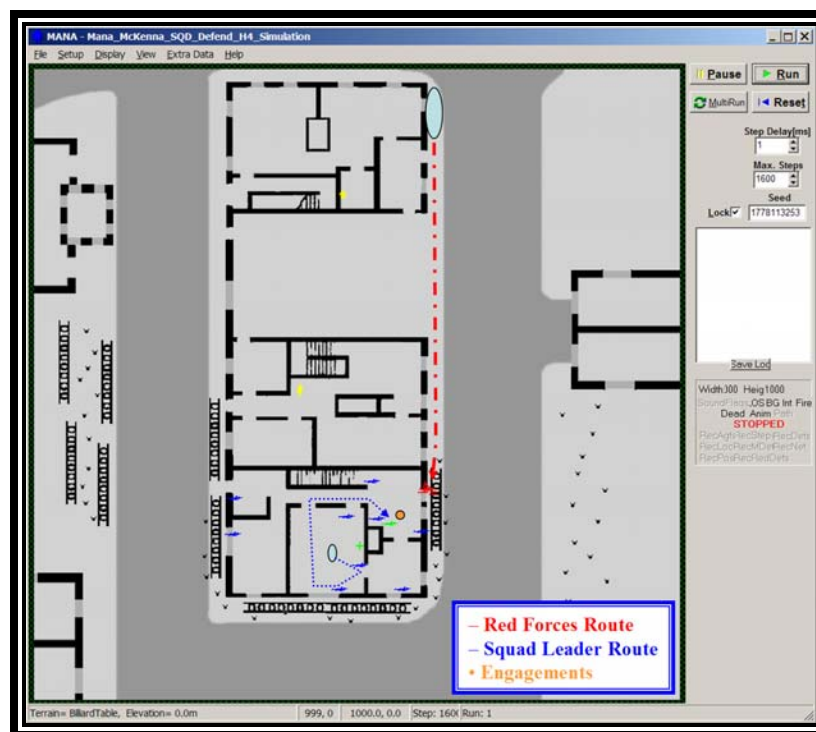


Figure F24. MANA Display of McKenna Defensive Scenario; End State, Blue Routes, and Engagement Locations for Squad Leader

The squad leader repositions one of his team leaders and an M249 SAW to the northeastern room (Figure F25). When the red force attempts to enter through the window, it becomes entangled in wire and is destroyed by the blue force's small-arms fire.

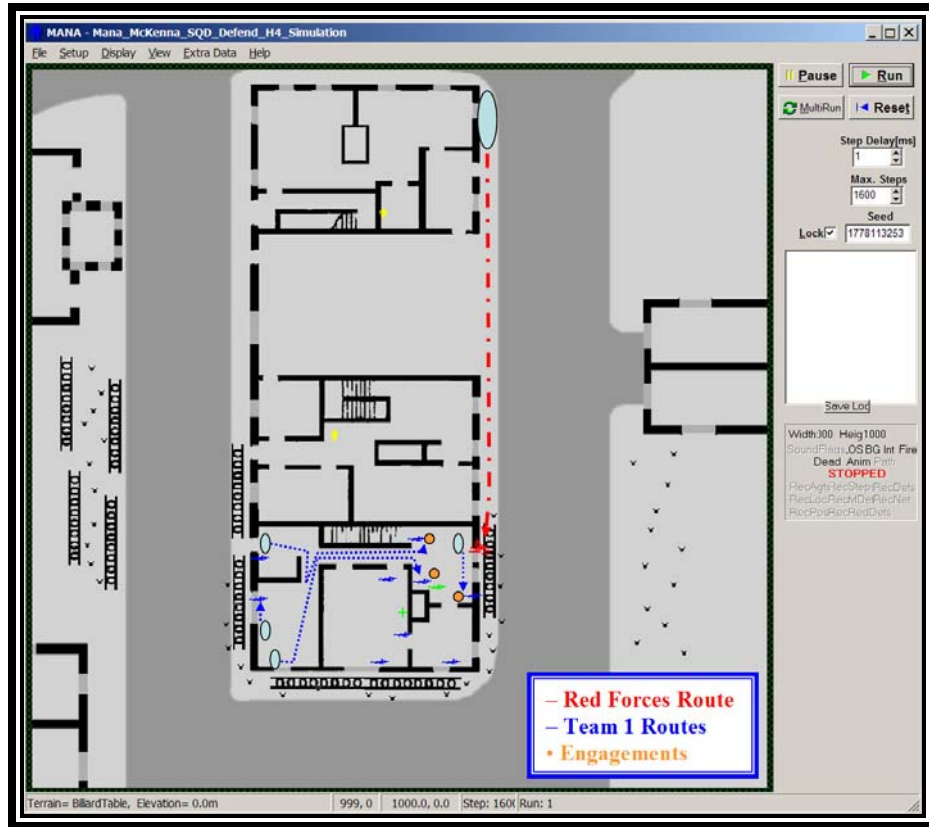


Figure F25. MANA Display of McKenna Defensive Scenario; End State, Blue Routes, and Engagement Locations for Team 1

This defensive scenario is the second scenario assessed during the data-collection phase of the study. As with the offensive scenarios, half the participants are informed the scenario is based on the 2002 MOUT experiment and half are told it is computer generated.

The XML file for this scenario is 2.188MB. The red force loses eight of its nine personnel. The blue squad and civilians sustain no loss. At near-real-time speed, the defensive scenario lasts nearly twenty-five seconds; at faster than real time, thirteen seconds.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX G. PARTICIPANT DEMOGRAPHICS, EXPERIENCE, AND TRAINING QUESTIONNAIRE

Participant Demographics, Experience, and Training Questionnaire

MOUT Evaluation/Validation Scenario Study

01 September 2003 through 31 December 2003

sponsored by the MOVES Institute, Naval Postgraduate School

Participant # _____

DEMOGRAPHICS

1. Name: _____
(Last) (First) (MI)
2. Age: _____ 3. Gender: ☐ Male ☐ Female
4. Service: ☐ Air Force ☐ Army ☐ Marine ☐ Navy ☐ DoD Civilian ☐ Other: _____
5. Component: ☐ Active Duty ☐ National Guard ☐ Reserve ☐ Not Applicable

EXPERIENCE

6. Highest Degree Completed: ☐ High School ☐ Associates Degree ☐ Bachelors ☐ Masters ☐ PhD
Area of concentration (History, Engineering, etc...)? _____
7. Military Rank: _____
8. Primary Branch (title & no.) _____
9. Functional Area or Secondary Branch (title & no.) _____
10. Prior Enlisted Service: ☐ No ☐ Yes, how long? Year(s): _____ Month(s): _____
11. How long have you been in the Army (to include enlisted service)? Year(s): _____ Month(s): _____
12. To what organization are you assigned (down to battalion level designation)? _____
13. Check the duty position(s) you have held
- | | | | |
|---|---|---|--|
| <input type="checkbox"/> Automatic Rifleman | <input type="checkbox"/> Grenadier | <input type="checkbox"/> Fire Team Leader | <input type="checkbox"/> Squad Leader |
| <input type="checkbox"/> Rifle Platoon Leader | <input type="checkbox"/> Scout Platoon Leader | <input type="checkbox"/> AT Platoon Leader | <input type="checkbox"/> Mortar Platoon Leader |
| <input type="checkbox"/> Rifle Company XO | <input type="checkbox"/> Rifle Company CDR | <input type="checkbox"/> Bn Staff (S3 shop) | <input type="checkbox"/> Other: _____ |
14. What is your current duty position? _____
15. How long has it been since you were in a line unit? Year(s): _____ Month(s): _____
16. To what type of line unit(s) were you assigned?
- | | | | |
|-----------------------------------|--------------------------------------|---|--|
| <input type="checkbox"/> Airborne | <input type="checkbox"/> Air Assault | <input type="checkbox"/> Light Infantry | <input type="checkbox"/> Mech Infantry |
| <input type="checkbox"/> Armor | <input type="checkbox"/> Cavalry | <input type="checkbox"/> Artillery | <input type="checkbox"/> Other: _____ |

17. Have you ever been deployed to a combat area or on a peacekeeping mission?
☐ No
☐ Yes, where, when, and type of unit? _____
18. Have you ever experienced combat?
☐ No
☐ Yes, where, when, and type of unit? _____
 What position did you occupy? _____
19. Have you ever conducted MOUT operations in a combat area or on a peacekeeping mission?
☐ No
☐ Yes, where, when, and type of unit? _____
20. In the past two years, how often did you play video games?
☐ 0 ☐ 1-10 hrs a year ☐ 1-10 hrs a month ☐ 3-4 hrs a week ☐ 4+ hrs a week
21. On average, how many hours a week do you currently play video games?
☐ 0 ☐ 0.5-1 hrs ☐ 1-2 hrs ☐ 2-3 hrs ☐ 3-4 hrs ☐ 4+ hrs
22. Have you ever played "first-person shooter" types of video games (e.g., Doom, Quake, Rainbow Six, Rogue Spear, Delta Force, Americas Army, etc.)?
☐ No
☐ Yes, list the games you recall playing _____
 What is your proficiency? ☐ Expert ☐ Average ☐ Novice
23. Have you ever used any combat models for training or studies (e.g., BCTP, Janus, VIC, COMBAT^{XXI}, JCATS, JSAF, ModSAF, OneSAF, etc.)?
☐ No
☐ Yes, list the models you have used _____
 How many times (days) have you used them? _____
 What was your impression of the model(s) portrayal of human behavior? _____
24. Have you ever participated in any MOUT studies or evaluations at the McKenna MOUT Site?
☐ No
☐ Yes, when? Month _____ Year _____
25. Have you ever participated in any MOUT studies, training, or evaluation at the JRTC MOUT Site?
☐ No
☐ Yes, when? Month _____ Year _____
26. Have you ever participated in any rotations at Grafenberg, JRTC, or NTC?
☐ Grafenberg # BLUEFOR Rotations _____ # OPFOR Rotations _____ # OC Rotations _____
☐ JRTC # BLUEFOR Rotations _____ # OPFOR Rotations _____ # OC Rotations _____
☐ NTC # BLUEFOR Rotations _____ # OPFOR Rotations _____ # OC Rotations _____

TRAINING

27. Have you ever received training in MOUT room clearing operations?

☐ No

☐ Yes, when and type of unit? _____

Where was the training conducted? _____

What TTPs were taught during this training? _____

How long was this training? Days _____ Hours _____

28. Have you ever received training at the McKenna MOUT Site?

☐ No

☐ Yes, when did you receive training? Month _____ Year _____

How long was this training? Days _____ Hours _____

THE END

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX H. CONSENT FORMS

Participants must complete three consent forms prior to participating in the study. What follows is a brief description of each form and example copies of the forms.

The Participant Consent Form consists of nine data elements and a signature section. The data elements outline the purpose of the study, who is conducting the study, the study procedures, risk & benefits to the participants, participant compensation for participating the study, confidentiality of participant information, the volunteer nature of the study, points of contact for the study, and a statement indicating the individual agrees to participate in the study once they sign the forms.

The Minimal Risk consent form outlines the type of materials the participant will be exposed to and ensures them they will not be intentionally harmed. The form provides a point of contact for the Naval Postgraduate School's Medical Monitor in the event the participant experiences some adverse physical ailment due to involvement with the study.

The last consent form is a Privacy Act Statement to ensure participants the Naval Postgraduate School will maintain all personal or reference material for this study. Research personnel will secure these materials in a locked facility.

Participant Consent Form

1. **Introduction.** You are invited to participate in a study of cognitive model validation procedures. With information gathered from you and other participants, we hope to discover insight on how to validate more effectively cognitive models for use in combat simulations. We ask you to read and sign this form indicating that you agree to be in the study. Please ask any questions you may have before signing.
2. **Background Information.** The Naval Postgraduate School's MOVES Institute is conducting this study.
3. **Procedures.** If you agree to participate in this study, the researcher will explain the tasks in detail. There will be two sessions: a) 45 minute in processing, training, and familiarization phase and b) 45 minute model validation phase where you will view five scenarios and establish the appropriate level of model validation. During the experiment you will be videotaped or filmed/photographed to help ensure accurate and comprehensive data collection.
4. **Risks and Benefits.** This research involves no risks or discomforts greater than those encountered during normal use of computer systems or classroom activities. The benefits to the participants are gaining techniques for assessing model performance and contributing to current research in cognitive model validation.
5. **Compensation.** No tangible reward will be given. A copy of the results will be available to you at the conclusion of the experiment.
6. **Confidentiality.** The records of this study will be kept confidential. No information will be publicly accessible which could identify you as a participant.
7. **Voluntary Nature of the Study.** If you agree to participate, you are free to withdraw from the study at any time without prejudice. You will be provided a copy of this form for your records.
8. **Points of Contact.** If you have any further questions or comments after the completion of the study, you may contact the research supervisor, Dr. Rudolph P. Darken (831) 656-7588 darken@nps.navy.mil.
9. **Statement of Consent.** I have read the above information. I have asked all questions and have had my questions answered. I agree to participate in this study.

Participant's Signature

Date

Researcher's Signature

Date



Minimal Risk Consent Form
MODELING, VIRTUAL ENVIRONMENTS, AND SIMULATIONS
ME Building, 700 Dyer Road
Naval Postgraduate School
Monterey, California 93943

Tel: 831-656-3733
DSN: 756-3733
Fax: 831-656-7599
srgoerge@nps.navy.mil

Participant: VOLUNTARY CONSENT TO BE A RESEARCH PARTICIPANT IN: Cognitive Model Validation Research

1. I have read, understand and been provided "Information for Participants" that provides the details of the below acknowledgments.
2. I understand that this project involves research. An explanation of the purposes of the research, a description of procedures to be used, identification of experimental procedures, and the extended duration of my participation have been provided to me.
3. I understand that this project does not involve more than minimal risk. I have been informed of any reasonably foreseeable risks or discomforts to me.
4. I have been informed of any benefits to me or to others that may reasonably be expected from the research.
5. All videotape, photographs, and written documents that could be used to identify me will be kept in a locked container.
6. I have been informed of any compensation and/or medical treatments available if injury occurs and is so, what they consist of, or where further information may be obtained.
7. I understand that my participation in this project is voluntary; refusal to participate will involve no penalty or loss of benefits to which I am otherwise entitled. I also understand that I may discontinue participation at any time without penalty or loss of benefits to which I am otherwise entitled.
8. I understand that the individual to contact should I need answers to pertinent questions about the research is Professor Rudy Darken, Principal Investigator, and about my rights as a research participant or concerning a research related injury is the Modeling Virtual Environments and Simulation Chairman. A full and responsive discussion of the elements of this project and my consent has taken place.

Medical Monitor: CAPT Nick Davenport, MC, USN; 656-7876; nadavenp@nps.navy.mil;
Flight Surgeon, Naval Postgraduate School

Signature of Principal Investigator

Date

Signature of Volunteer

Date

Signature of Witness

Date



Privacy Act Statement

MODELING, VIRTUAL ENVIRONMENTS, AND SIMULATIONS

ME Building, 700 Dyer Road
Naval Postgraduate School
Monterey, California 93943

Tel: 831-656-3733

DSN: 756-3733

Fax: 831-656-7599

srgoerge@nps.navy.mil

1. Purpose: Cognitive model validation data will be collected to enhance knowledge, and to develop processes to improve the development and validation of cognitive models for use in combat simulations.
2. Use: Cognitive model validation data will be used for statistical analysis by the Departments of the Navy and Defense, and other U.S. Government agencies, provided this use is compatible with the purpose for which the information was collected. The Naval Postgraduate School in accordance with the provisions of the Freedom of Information Act may grant use of the information to legitimate non-government agencies or individuals.
3. Disclosure/Confidentiality:
 - a. I have been assured that my privacy will be safeguarded. I will be assigned a control or code number, which thereafter will be the only identifying entry on any of the research records. The Principal Investigator will maintain the cross-reference between name and control number. It will be decoded only when beneficial to me or if some circumstances, which is not apparent at this time, would make it clear that decoding would enhance the value of the research data. In all cases, the provisions of the Privacy Act Statement will be honored.
 - b. I understand that a record of the information contained in this Consent Statement or derived from the experiment described herein will be retained permanently in a locked container at the Naval Postgraduate School or by higher authority. I voluntarily agree to its disclosure to agencies or individuals indicated in paragraph 3 and I have been informed that failure to agree to such disclosure may negate the purpose for which the experiment was conducted.
 - c. I also understand that disclosure of the requested information, including my Social Security Number, is voluntary.

Signature of Volunteer	Name, Grade/Rank (if applicable)	DOB	SSN	Date
------------------------	----------------------------------	-----	-----	------

APPENDIX I. DEBRIEFING HANDOUT



MODELING, VIRTUAL ENVIRONMENTS, AND SIMULATIONS

ME Building, 700 Dyer Road
Naval Postgraduate School
Monterey, California 93943

Tel: 831-656-3733
DSN: 756-3733
Fax: 831-656-7599
srgoerge@nps.navy.mil

The use of virtual environments in training and education has been an expanding field for the last two decades. With recent developments in psychology and models & simulations, cognitive (human behavior) models are asked to do more and are increasing in complexity. In order to insure these models are meeting the needs they were designed for and to better understand their strengths and weaknesses when are researching methods for validating the capabilities of these models.

The study you have just completed is concerned with gathering information on how individuals evaluate human and/or validate simulated human behavior. You spent a session assessing behaviors for a series of ground combat scenarios.

Four separate groups are being examined in order to determine potential user bias using current validation procedures for assessing human and simulated behaviors. All groups were exposed to a computer simulation or replication of a computer simulation demonstrating human behaviors. One group assessed real human behaviors. A second group assessed the validity of simulated behaviors. The last two groups assessed both real human and simulated behaviors.

The research personnel observed and recorded information concerning the perceptions of the participants. This data will be used for the redesign and implementation of more useful validation procedures for cognitive models. Your assistance in this project will contribute to the production of more useful virtual environments that provide users with more realistic human entities to interact with as friendly, neutral, and/or enemy forces.

If you have any questions about the study, please ask your research assistant. Until 30 September 2003, please do not discuss this experiment with anyone except our research personnel to prevent influencing any future participants. Thank you for your participation in this study.

The research supervisor, MAJ Simon R. Goerger, for this study can be contacted at (831) 656-3733 or Email: srgoerge@ nps.navy.mil.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX J. EXPERIMENT EXIT QUESTIONNAIRE

The following is an example of the Exit Questionnaire provided to participants at the end of the study. Participants are asked to take the form with them, complete it at their leisure, and return it to a study drop box once it is complete. Participants are not required to complete this form.

MOUT Evaluation/Validation Scenario Study
22 October 2003 through 30 October 2003
Sponsored by the MOVES Institute, Naval Postgraduate School
Exit Questionnaire

Participant # _____ Date _____

Please answer all questions at the completion of the study. All information will be kept confidential.

1. Do you feel that time pressure had an impact upon your ability to make a complete assessment?

☐ No

☐ Yes, in what way? _____

Other Comments: _____

2. Did you know enough about the model?

☐ No, why? _____

☐ Yes

Other Comments: _____

3. Did you know enough about urban operations to assess performance during these scenarios?

☐ No, why? _____

☐ Yes

Other Comments: _____

4. Did you know enough about the MOUT tactics, techniques, and procedures to assess performance during these scenarios?

☐ No, why? _____

☐ Yes

Other Comments: _____

5. Did you feel the assessment sheets were adequate to allow you to assess performance during these scenarios?

☐ No, why? _____

☐ Yes

Other Comments: _____

6. In general, how realistic were the MOUT scenario(s)?

- ☐ Very Realistic
- ☐ Somewhat Realistic
- ☐ Unsure
- ☐ Somewhat Unrealistic
- ☐ Very Unrealistic

7. Was there any single task that you considered more difficult to assess than the others?

- ☐ No
- ☐ Yes, which one and why? _____

Other Comments: _____

8. Please list any difficulties that were not previously mentioned.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX K. ASSESSMENT WORKSHEETS

For this research, participants use a series of modified assessment forms based on *FM 7-8: Infantry Rifle Platoon and Squad*, 2001 and *ARTEP 7-8-MTP: Mission Training Plan for the Infantry Rifle Platoon and Squad*, 2001 [DEPA 01g] [ARTP 01]. For each of the studies, participants were asked to assess the same questions. The assessment forms used varied only in the assessment scales.

The following are five sets example sheets used for the task assessment. Each example uses a different assessment scale, the 7-Point Likert Scale, 5-Point Likert Scale, or the Go/No-Go Scale. Example number one is a complete assessment form for the task “React to Sniper” to demonstrate how the MTP work sheets are modified for the validation process. The next two examples are the first page of the 5-Point Likert Scale and Go/No-Go Scale assessment worksheets for the task “React to Sniper” to demonstrate how the worksheets differ from the 7-Point Likert Scale. The final two assessment worksheet examples show the assessment form used for the Scenario Level and Overall Assessment of the model based on the 7-Point Likert Scale.

1. OFFENSIVE SCENARIO, TASK: REACT TO SNIPERS (INFANTRY SQUAD) (07-3-1406) 7-POINT LIKERT SCALE

PARTICIPANT ID#: _____

REFERENCE(S): (FM 21-60) (FM 24-35) (FM 24-35-1) (FM 7-4 (3-21.94)) (FM 7-5 (3-21.9)) (FM 7-7) (FM 7-7J) (FM 7-8) (FM 7-85) (FM 7-92) (FM 90-10(HTF)) (FM 90-10-1)

CONDITION: The squad is conducting operations as part of a larger force and receives fire from an enemy sniper. The squad must react immediately for their protection. All necessary personnel and equipment are available. The squad has communications with higher, adjacent, and subordinate elements. The squad has been provided guidance on the rules of engagement (ROE) and/or rules of interaction (ROI). Coalition forces and noncombatants may be present in the operational environment.

TASK STANDARD: The squad reacts to the sniper in accordance with (IAW) tactical standing operating procedures (TSOP), the order, and/or commander's guidance. The squad correctly locates and then bypasses, eliminates, or forces the withdrawal of the enemy sniper while disengaging the element in the kill zone. The squad complies with the ROE and/or ROI.

ASSESSMENT SCALE: Use the following scale to assess performance of the squad as it performs this task.

- 1 – Strongly agree** the task, step, or performance measure was *improperly* performed
- 2 – Agree** the task, step, or performance measure was *improperly* performed
- 3 – Not sure** but tend to agree the task, step, or performance measure was *improperly* performed
- 4 – Undecided**
- 5 – Not sure** but tend to agree the task, step, or performance measure was *properly* performed
- 6 – Agree** the task, step, or performance measure was *properly* performed
- 7 – Strongly agree** the task, step, or performance measure was *properly* performed
- NA - Not applicable** or no means of determining

TASK(S), STEP(S) and PERFORMANCE MEASURE(S)	ASSESSMENT
1. Squad conducts actions on contact (sniper fire). <i>Comments:</i>	1-2-3-4-5-6-7 NA

TASK(S), STEP(S) and PERFORMANCE MEASURE(S)	ASSESSMENT
<p>a. Returns fire immediately to destroy or suppress the enemy. <i>Comments:</i></p>	<p>1-2-3-4-5-6-7 NA</p>
<p>b. Deploys to covered and concealed positions, if available. <i>Comments:</i></p>	<p>1-2-3-4-5-6-7 NA</p>
<p>c. Conducts battle drills, as necessary. <i>Comments:</i></p>	<p>1-2-3-4-5-6-7 NA</p>
<p>d. Maintains visual contact with the enemy while continuing to develop the situation through reconnaissance or surveillance. <i>Comments:</i></p>	<p>1-2-3-4-5-6-7 NA</p>
<p>e. Maintains cross talk with all squad elements using FBCB2, FM, or other tactical means. <i>Comments:</i></p>	<p>1-2-3-4-5-6-7 NA</p>
<p>2. Squad reacts to enemy sniper fire. <i>Comments:</i></p>	<p>1-2-3-4-5-6-7 NA</p>

TASK(S), STEP(S) and PERFORMANCE MEASURE(S)	ASSESSMENT
<p>a. Reports contact to higher headquarters using FBCB2, FM, or other tactical means. Comments:</p>	<p>1-2-3-4-5-6-7 NA</p>
<p>b. <i>(If the sniper is killed, go to sub task 2.c)</i> Bypasses the sniper. Comments:</p>	<p>1-2-3-4-5-6-7 NA</p>
<p>(1) The squad uses smoke to obscure the enemy sniper's view. Comments:</p>	<p>1-2-3-4-5-6-7 NA</p>
<p>(2) The squad uses available fires to suppress the sniper. Comments:</p>	<p>1-2-3-4-5-6-7 NA</p>
<p>(3) The squad maneuvers to break contact with the sniper. Note. The squad leader may choose to call for indirect fire on the sniper position. Comments:</p>	<p>1-2-3-4-5-6-7 NA</p>
<p>c. <i>(If the sniper is NOT killed, go to sub task 3)</i> Eliminates the sniper. Comments:</p>	<p>1-2-3-4-5-6-7 NA</p>

TASK(S), STEP(S) and PERFORMANCE MEASURE(S)	ASSESSMENT
(1) Complies with ROE and/or ROI. <i>Comments:</i>	1-2-3-4-5-6-7 NA
(2) The squad uses smoke to obscure the enemy sniper's view. <i>Comments:</i>	1-2-3-4-5-6-7 NA
(3) The squad uses available firepower to suppress and fix the sniper. <i>Comments:</i>	1-2-3-4-5-6-7 NA
(4) The squad maneuvers to close with the sniper and eliminate or force him to withdraw. <i>Comments:</i>	1-2-3-4-5-6-7 NA
3. Squad consolidates and reorganizes as necessary. <i>Comments:</i>	1-2-3-4-5-6-7 NA
4. Squad treats and evacuates casualties as necessary. <i>Comments:</i>	1-2-3-4-5-6-7 NA

TASK(S), STEP(S) and PERFORMANCE MEASURE(S)	ASSESSMENT
5. Squad secures enemy prisoners of war (EPW), if applicable. <i>Comments:</i>	1-2-3-4-5-6-7 NA
6. Squad processes captured documents and/or equipment, if applicable. <i>Comments:</i>	1-2-3-4-5-6-7 NA
7. * Squad leader reports to higher headquarters as required using FBCB2, FM, or other tactical means. <i>Comments:</i>	1-2-3-4-5-6-7 NA
8. Squad continues operations as directed. <i>Comments:</i>	1-2-3-4-5-6-7 NA

NOTE * Indicates a leader task.

TASK PERFORMANCE SUMMARY BLOCK								
ASSESSMENT CATEGORY	1	2	3	4	5	6	7	NA
Total tasks, steps, and performance measures evaluated at this level.								

TASK: React to Snipers (Infantry Squad) (07-3-1406)	ASSESSMENT
<p>Squad performances this task to standard.</p> <p><i>Comments:</i></p>	<p>1-2-3-4-5-6-7</p> <p>NA</p>

2. OFFENSIVE SCENARIO, TASK: REACT TO SNIPERS (INFANTRY SQUAD) (07-3-1406) 5-POINT LIKERT SCALE

PARTICIPANT ID#:_____

REFERENCE(S): (FM 21-60) (FM 24-35) (FM 24-35-1) (FM 7-4 (3-21.94)) (FM 7-5 (3-21.9)) (FM 7-7) (FM 7-7J) (FM 7-8) (FM 7-85) (FM 7-92) (FM 90-10(HTF)) (FM 90-10-1)

CONDITION: The squad is conducting operations as part of a larger force and receives fire from an enemy sniper. The squad must react immediately for their protection. All necessary personnel and equipment are available. The squad has communications with higher, adjacent, and subordinate elements. The squad has been provided guidance on the rules of engagement (ROE) and/or rules of interaction (ROI). Coalition forces and noncombatants may be present in the operational environment.

TASK STANDARD: The squad reacts to the sniper in accordance with (IAW) tactical standing operating procedures (TSOP), the order, and/or commander's guidance. The squad correctly locates and then bypasses, eliminates, or forces the withdrawal of the enemy sniper while disengaging the element in the kill zone. The squad complies with the ROE and/or ROI.

ASSESSMENT SCALE: Use the following scale to assess performance of the squad as it performs this task.

- 1 – Strongly agree** the task, step, or performance measure was *improperly* performed
- 2 – Agree** the task, step, or performance measure was *improperly* performed
- 3 – Undecided**
- 4 – Agree** the task, step, or performance measure was *properly* performed
- 5 – Strongly agree** the task, step, or performance measure was *properly* performed
- NA - Not applicable** or no means of determining

TASK(S), STEP(S) and PERFORMANCE MEASURE(S)	ASSESSMENT
1. Squad conducts actions on contact (sniper fire). <i>Comments:</i>	1-2-3-4-5 NA

3. OFFENSIVE SCENARIO, TASK: REACT TO SNIPERS (INFANTRY SQUAD) (07-3-1406) GO/NO-GO SCALE

PARTICIPANT ID#:_____

REFERENCE(S): (FM 21-60) (FM 24-35) (FM 24-35-1) (FM 7-4 (3-21.94)) (FM 7-5 (3-21.9)) (FM 7-7) (FM 7-7J) (FM 7-8) (FM 7-85) (FM 7-92) (FM 90-10(HTF)) (FM 90-10-1)

CONDITION: The squad is conducting operations as part of a larger force and receives fire from an enemy sniper. The squad must react immediately for their protection. All necessary personnel and equipment are available. The squad has communications with higher, adjacent, and subordinate elements. The squad has been provided guidance on the rules of engagement (ROE) and/or rules of interaction (ROI). Coalition forces and noncombatants may be present in the operational environment.

TASK STANDARD: The squad reacts to the sniper in accordance with (IAW) tactical standing operating procedures (TSOP), the order, and/or commander's guidance. The squad correctly locates and then bypasses, eliminates, or forces the withdrawal of the enemy sniper while disengaging the element in the kill zone. The squad complies with the ROE and/or ROI.

ASSESSMENT SCALE: Use the following scale to assess performance of the squad as it performs this task.

Go – The task, step, or performance measure was properly performed

No Go – The task, step, or performance measure was **NOT** properly performed

T - Trained. The unit successfully performed all subtasks.

P - Needs Practice. The unit needs to practice the task. All critical subtasks were performed successfully, but one or more noncritical subtasks were performed unsuccessfully.

U - Untrained. The unit incorrectly performed or failed to perform one or more critical subtasks.

NA - **Not applicable** or no means of determining

TASK(S), STEP(S) and PERFORMANCE MEASURE(S)	ASSESSMENT
1. Squad conducts actions on contact (sniper fire). <i>Comments:</i>	Go - No Go NA

4. OFFENSIVE SCENARIO #1, TASK: ASSESSMENT SUMMARY 7-POINT LIKERT SCALE

PARTICIPANT ID#: _____

REFERENCE(S): (FM 7-5 (3-21.9)) (FM 7-7J) (FM 7-8) (FM 90-10(HTF)) (FM 90-10-1)

CONDITION: The squad is conducting operations as part of a larger force in an urban environment and has received an operation order (OPORD) or fragmentary order (FRAGO) to assault and clear a building. The building has two levels and contains a squad-sized enemy element. All necessary personnel and equipment are available. The squad has communications with higher, adjacent, and subordinate elements. The squad has been provided guidance on the Rules of Engagement (ROE) and/or Rules of Interaction (ROI). Coalition forces and noncombatants may be present in the operational environment.

TASK STANDARD: The squad assaults and clears the building in accordance with (IAW) tactical standing operating procedures (TSOP), the order, and/or commander's guidance. The squad kills, captures, or forces the withdrawal of all enemy in the building. The squad complies with the ROE and/or ROI.

ASSESSMENT SCALE: Use the following scale to assess performance of the squad as it performs this task.

- 1 – **Strongly agree** the task, step, or performance measure was *improperly* performed
- 2 – **Agree** the task, step, or performance measure was *improperly* performed
- 3 – **Not sure** but tend to agree the task, step, or performance measure was *improperly* performed
- 4 – **Undecided**
- 5 – **Not sure** but tend to agree the task, step, or performance measure was *properly* performed
- 6 – **Agree** the task, step, or performance measure was *properly* performed
- 7 – **Strongly agree** the task, step, or performance measure was *properly* performed
- NA - **Not applicable** or no means of determining

TASK(S), STEP(S) and PERFORMANCE MEASURE(S)	ASSESSMENT
1. Conduct Tactical Movement in a Built-up Area (Infantry Squad) (07-3-1279) <i>Comments:</i>	1-2-3-4-5-6-7 NA

TASK(S), STEP(S) and PERFORMANCE MEASURE(S)	ASSESSMENT
2. React to Snipers (Infantry Squad) (07-3-1406) <i>Comments:</i>	1-2-3-4-5-6-7 NA
3. Conduct Tactical Movement in a Built-up Area (Infantry Squad) (07-3-1279) <i>Comments:</i>	1-2-3-4-5-6-7 NA

TASK PERFORMANCE SUMMARY BLOCK								
ASSESSMENT CATEGORY	1	2	3	4	5	6	7	NA
Total tasks, steps, and performance measures evaluated at this level.								

Offensive Scenario #1:	ASSESSMENT
This scenario was executed by: <input type="checkbox"/> Simulated Behaviors <input type="checkbox"/> Scripted Human Performance	
Squad performances this tasks to standard. <i>Comments:</i>	1-2-3-4-5-6-7 NA

5. URBAN OPERATIONS ASSESSMENT SUMMARY

PARTICIPANT ID#: _____

REFERENCE(S): (FM 7-5 (3-21.9)) (FM 7-7) (FM 7-7J) (FM 7-8)

CONDITION: Given a human behavior representation model and the performance of a light infantry squad, assess the model's ability to portray a squad is conducting operations as part of a larger force in urban terrain. The squad must perform both defensive and offensive operations. All necessary personnel and equipment are available. The model portrays squad communications with higher, adjacent, and subordinate elements. The simulated and real world squads have been provided guidance on the rules of engagement (ROE) and/or rules of interaction (ROI). The scenarios portray coalition forces and noncombatants which may be present in the operational environment.

TASK STANDARD: The squad/model defends and assaults in accordance with tactical standing operating procedures (TSOP), the order, and/or commander's guidance. The squad/model deploys and moves similar to US forces operating in an urban environment in accordance with current US Military tactics techniques and procedures as prescribed in appropriate field manuals and soldier skill manuals. The squad/model destroys or defeats the enemy force within its area. The squad/model complies with the ROE and/or ROI.

ASSESSMENT SCALE: Use the following scale to assess performance of the squad as it performs this task.

- 1 – **Strongly agree** the task, step, or performance measure was *improperly* performed
- 2 – **Agree** the task, step, or performance measure was *improperly* performed
- 3 – **Not sure** but tend to agree the task, step, or performance measure was *improperly* performed
- 4 – **Undecided**
- 5 – **Not sure** but tend to agree the task, step, or performance measure was *properly* performed
- 6 – **Agree** the task, step, or performance measure was *properly* performed
- 7 – **Strongly agree** the task, step, or performance measure was *properly* performed
- NA - **Not applicable** or no means of determining

TASK(S), STEP(S) and PERFORMANCE MEASURE(S)	ASSESSMENT
1. Practice Defensive Scenario #1 was: <input type="checkbox"/> Simulated Behaviors <input type="checkbox"/> Scripted Human Performance	
2. Practice Defensive Scenario #1 <i>Comments:</i>	1-2-3-4-5-6-7 NA
3. Offensive Scenario #1 was: <input type="checkbox"/> Simulated Behaviors <input type="checkbox"/> Scripted Human Performance	
4. Offensive Scenario #1 <i>Comments:</i>	1-2-3-4-5-6-7 NA
5. Defensive Scenario #1 was: <input type="checkbox"/> Simulated Behaviors <input type="checkbox"/> Scripted Human Performance	
6. Defensive Scenario #1 <i>Comments:</i>	1-2-3-4-5-6-7 NA
7. Offensive Scenario #2 was: <input type="checkbox"/> Simulated Behaviors <input type="checkbox"/> Scripted Human Performance	
8. Offensive Scenario #2 <i>Comments:</i>	1-2-3-4-5-6-7 NA

TASK PERFORMANCE SUMMARY BLOCK								
ASSESSMENT CATEGORY	1	2	3	4	5	6	7	NA
Total tasks, steps, and performance measures evaluated at this level.								

General Performance:	ASSESSMENT
<p>At the individual squad member level, the squad/model executed the scenario(s) in a manner appropriate for US forces operating in urban terrain.</p> <p><i>Comments:</i></p>	<p>1-2-3-4-5-6-7 NA</p>
<p>At the squad level, the squad/model executed the scenario(s) in a manner appropriate for US forces operating in urban terrain.</p> <p><i>Comments:</i></p>	<p>1-2-3-4-5-6-7 NA</p>

General Performance:	ASSESSMENT
<p>I am confident the squad/model can execute individual tasks in similar scenario(s) and environments in a manner appropriate for US forces operating in urban terrain.</p> <p><i>Comments:</i></p>	<p>1-2-3-4-5-6-7 NA</p>
<p>I am confident the squad/model can execute squad level tasks in similar scenario(s) and environments in a manner appropriate for US forces operating in urban terrain.</p> <p><i>Comments:</i></p>	<p>1-2-3-4-5-6-7 NA</p>

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX L. BRIEFING SCRIPTS

Appendix L. consists of eight briefing scripts: In-Brief, Assessment Procedure Brief, Model Familiarization Brief, Practical Exercise Brief, Control Group Brief, Study Group 1 Brief, Study Groups 2 & 3 Brief, Assessment Study Brief, and the Debrief. Each participant receives the In-Brief and Debrief. Research personnel use the briefing scripts for the introduction and general data collection portion of the study. The scripts for the training and warm-up phases of the study are imbedded in the slide presentations used to facilitate the study (Appendix M. Experiment Briefing Slides).

1. IN-BRIEF

Welcome, my name is _____. Thank you for your assistance with today's experiment. Today's experiment is conducted by the Naval Postgraduate School's MOVES Institute and is concerned with procedures for assessing the performance of cognitive models.

This experiment is not a test of your intelligence or performance. Rather, it is an evaluation of cognitive (human behavior) assessment procedures. *(For Military Personnel) Your performance will not be recorded in your personnel records but is intended for research.* All information collected is for academic research only. Before starting the experiment, you will be asked to read and sign a series of consent forms. Upon signing the consent forms, you will undergo a thirty minute model familiarization and validation procedure train-up prior to your exposure to a series of simulation scenarios where you will be asked to using specified procedures to assess the validity of the behaviors being portrayed. Upon completing the scenarios, you will undergo a short debriefing. If there are no questions, please read and sign the Participant Consent Form.

After signing your Participant Consent Form, please read and sign the Minimal Risk Consent Form. There are two copies of this form. One copy is for you to take with you and the other remains with your packet for the study records. After signing your Minimal Risk Consent Forms, have the person sitting at the desk next to you sign the witness block on the form.

The final form is the Privacy Act Statement. Please read and sign this form.

Pass out NEO Five Factor Inventory

The final portion of the in-processing is the NEO Five Factor Inventory. This is a four-page booklet with sixty questions designed to categorize your personality. **Do not** write in this booklet until told to do so. You have five possible answers, Strongly Disagree (SD), Disagree (D), Neutral (N), Agree (A), and Strongly Agree (SA). You provide the answer to each question by filling in the answer space on the third page of the booklet. Write only where indicated in the booklet. At this time read the directions on the first page of the booklet on your own. When done, turn your booklet upside down. Ready... Begin.

Open the booklet to the second page. Complete the header section, placing your name, age, gender, and today's date at the top. Today's date is _____. Place your pens down when you have completed this task.

A reminder that you are only to write your answers to the questions on the third page of the booklet. Please complete all sixty questions plus the three final questions at the bottom of page three. Your answers will be reviewed and a one-page summary categorizing your personality will be provided to you upon completion of the data collection phase. Are there any questions? At this time, you will have ten minutes to complete this questioner. When finished, turn your test upside down. Ready... Begin.

Pick-up consent forms and place down assessment forms

This completes the in-processing portion of this study.

2. ASSESSMENT PROCEDURE BRIEF

For the next ten minutes, you will undergo a series of instruction on assessment procedures. At the end of this instruction, you will undergo a practical exercise to demonstrate your understanding of the procedures taught today. I want you to use these procedures to assess performance during the data collection phase of this study.

To your front is a project display of the procedures you will be asked to complete for the data collection portion of this experiment...

(Conduct Assessment Procedure Brief using Introduction.ppt)

3. MODEL FAMILIARIZATION BRIEF

Prior to beginning the study, you will undergo a five-minute model familiarization phase. This is to help you become comfortable with the look, feel, and capabilities of the model prior to starting the experiment.

To your front is a project display of the MANA interface...

(Conduct Model Familiarization Phase using Introduction.ppt)

4. PRACTICAL EXERCISE BRIEF

You have just completed the assessment procedure and model familiarization instruction. You will now be given an opportunity to demonstrate your knowledge of these procedures by assessing the performance of a human behavior model executing a short scenario.

Defensive Scenario 01, Task 1:

Task: “Your first task of the defensive scenario is to depict the locations of forces as you have placed them to defend the first floor of building H4.”

Condition: “Given a classroom environment, data collection work sheets, scratch papers, maps of the scenario area, writing utensils, and a ground combat simulation running a **computer/soldier** generated scenario consisting of a dismounted infantry squad conducting operations in an urban environment, assess the behaviors presented on the screen. During the scenario, the model will be paused to allow you time to record your observations. Reminder the higher the value used (1-7 scale) the better the performance.

Standards: “Assess the behaviors utilizing the worksheets provided, your knowledge of urban operations, FM 90-10-1: Combined Arms Operations in Urban Terrain, ARTEP 7-8-MTP: Mission Training Plan for the Infantry Rifle Platoon and Squad, FM 101-5-1: Operational Terms and Graphics, and the assessment procedures you have been taught just prior to this exercise.”

“You have three minutes. Ready,... Begin.”

Defensive Scenario 01, Task 1 Complete

Record Time (to secs)

During scenario assessment, I will fast forward through the scenario one time. I will then play the scenario a second time through at near real time speeds. In order to facilitate your ability to record information on the Data Collection Sheets, I will occasionally pause the scenario after the execution of selected tasks and ask you to record your observations.

Record Time (to secs)

Defensive Scenario 01, Task 2:

Task: “Your second task of the defensive scenario is to assess the individual and squad behaviors as they ‘Conduct a Strongpoint Defense of a Building’ (Infantry Squad) (07-3-1162). Conditions and standards are unchanged.”

“You have three minutes. Ready,... Begin.”

Defensive Scenario 01, Task 2 Complete

Record Time (to secs)

Do you have any questions before we begin?

5. CONTROL GROUP BRIEF

For the remainder of the study you will assess the performance of soldiers performing urban combat operations in three separate scenarios in order to determine their level of proficiency. You will assess the performance of individual subtasks and the overall performance of the squad. On the screen is a video example of performance that will be played back using the MANA interface.

Show Video Clip of McKenna Squad MOUT Study.

During each scenario assessment, I will fast-forward each scenario one time through. I will then play each scenario a second time through at near real time speeds. In order to facilitate your ability to record information on the Data Collection Sheets, I will occasionally pause the scenario after the execution of selected tasks and ask you to record your observations.

Do you have any questions before we begin?

6. STUDY GROUP 1 BRIEF

For the remainder of the study you will assess the performance of cognitive model representing human performance during urban combat operations in three separate scenarios in order to determine the level of validity of the model. You will assess the performance of individual subtasks and the overall performance of the model.

During each scenario assessment, I will fast-forward each scenario one time through. I will then play each scenario a second time through at near real time speeds. In order to facilitate your ability to record information on the Data Collection Sheets, I will occasionally pause the scenario after the execution of selected tasks and ask you to record your observations.

Do you have any questions before we begin?

7. STUDY GROUPS 2 & 3 BRIEF

For the remainder of the study you will assess the performance of cognitive model representing human performance and real soldiers' performance during urban combat operations in three separate scenarios in order to determine the level of validity of the model. You will assess the performance of individual subtasks and the overall

performance of the model or soldiers. On the screen is a video example of performance, which will be played back, using the MANA interface for the **first and second scenario/ third scenario**.

Show Video Clip of McKenna Squad MOUT Study.

During each scenario assessment, I will fast-forward each scenario one time through. I will then play each scenario a second time through at near real time speeds. During this assessment, I will manipulate the scenario through a series of situations. In order to facilitate your ability to record information on the Data Collection Sheets, I will occasionally pause the scenario after the execution of selected tasks and ask you to record your observations.

Do you have any questions before we begin?

8. ASSESSMENT STUDY BRIEF

“Before of you is a projection of a computer simulation with human behaviors generated by (**real world data/computer program**) which you will be observing for your scenario assessment(s) today. On your desk are a blue pen and a Participant Record Folder. The Participant Record Folder contains Assessment Sheet(s), Participant Map(s) of the scenario areas of interest, and scratch paper. On the walls are posted model interface and assessment worksheet posters. These materials will be utilized for the assessment of a *cognitive model or soldier* performance during urban operations in a ground combat domain.

You will be assessing one defensive and two offensive scenarios.

From now until completion of the study, do not interact with **anyone**. Any interaction or verbal reflection may influence the assessment of other participants and bias the results of the study. Before you begin, do you have any questions?”

Operations PPT presentation (tactical overview, ROE, and scenario)

Start Study Timer

Offensive Scenario 01, Task 1:

Task: “Your first task of the first offensive scenario is to draw the route you would have the attacking squad take to building H4.”

Condition: “Given a classroom environment, data collection work sheets, scratch papers, maps of the scenario area, writing utensils, and a ground combat simulation running a **computer/soldier** generated scenario consisting of a dismounted infantry squad conducting operations in an urban environment, assess the behaviors presented on the screen. During the scenario, the model will be paused to allow you time to record your observations. Reminder the higher the value used (1-7 scale) the better the performance.

Standards: “Assess the behaviors utilizing the worksheets provided, your knowledge of urban operations, FM 90-10-1: Combined Arms Operations in Urban Terrain, ARTEP 7-8-MTP: Mission Training Plan for the Infantry Rifle Platoon and Squad, FM 101-5-1: Operational Terms and Graphics, and the assessment procedures you have been taught just prior to this exercise.”

“You have three minutes. Ready,... Begin.”

Offensive Scenario 01, Task 1 Complete

Record Time (to secs)

During scenario assessment, I will fast forward through the scenario one time. I will then play the scenario a second time through at near real time speeds. In order to facilitate your ability to record information on the Data Collection Sheets, I will occasionally pause the scenario after the execution of selected tasks and ask you to record your observations.

Record Time (to secs)

Offensive Scenario 01, Task 2:

Task: “Your second task is to assess the individual and squad behaviors as they ‘Conduct Tactical Movement in a Built-up Area’ (Infantry Squad) (07-3-1279). Conditions and standards are unchanged.”

“You have three minutes. Ready,... Begin.”

Offensive Scenario 01, Task 2 Complete

Record Time (to secs)

Offensive Scenario 01, Task 3.

Task: “Your third task is to assess the individual and squad behaviors as they ‘React to Snipers’ (Infantry Squad) (07-3-1406). Conditions and standards are unchanged.”

“You have three minutes. Ready,... Begin.”

Offensive Scenario 01, Task 3 Complete

Record Time (to secs)

Offensive Scenario 01, Task 4.

Task: “Your forth task is to assess the individual and squad behaviors as they ‘Conduct Tactical Movement in a Built-up Area’ (Infantry Squad) (07-3-1279). Conditions and standards are unchanged.”

“You have three minutes. Ready,... Begin.”

Offensive Scenario 01, Task 4 Complete

Record Time (to secs)

Offensive Scenario 01, Task 5.

Task: “Your fifth task is to assess the overall individual and squad behaviors for this scenario utilizing the last three tasks and other observations you have made. Conditions and standards are unchanged.”

“You have three minutes. Ready,... Begin.”

Offensive Scenario 01, Task 5 Complete

Record Time (to secs)

Next, you will assess a defensive scenario.

Defensive Scenario 01, Task 1:

Task: “Your first task of the defensive scenario is to depict the locations of forces as you have placed them to defend the first floor of building H4.”

Condition: “Given a classroom environment, data collection work sheets, scratch papers, maps of the scenario area, writing utensils, and a ground combat simulation running a **computer/soldier** generated scenario consisting of a dismounted infantry squad conducting operations in an urban environment, assess the behaviors presented on the screen. During the scenario, the model will be paused to allow you time to record your observations. Reminder the higher the value used (1-7 scale) the better the performance.

Standards: “Assess the behaviors utilizing the worksheets provided, your knowledge of urban operations, FM 90-10-1: Combined Arms Operations in Urban Terrain, ARTEP 7-8-MTP: Mission Training Plan for the Infantry Rifle Platoon and Squad, FM 101-5-1: Operational Terms and Graphics, and the assessment procedures you have been taught just prior to this exercise.”

“You have three minutes. Ready,... Begin.”

Defensive Scenario 01, Task 1 Complete

Record Time (to secs)

During scenario assessment, I will fast forward through the scenario one time. I will then play the scenario a second time through at near real time speeds. In order to facilitate your ability to record information on the Data Collection Sheets, I will occasionally pause the scenario after the execution of selected tasks and ask you to record your observations.

Record Time (to secs)

Defensive Scenario 01, Task 2:

Task: “Your second task of the defensive scenario is to assess the individual and squad behaviors as they ‘Conduct a Strongpoint Defense of a Building’ (Infantry Squad) (07-3-1162). Conditions and standards are unchanged.”

“You have three minutes. Ready,... Begin.”

Defensive Scenario 01, Task 2 Complete

Record Time (to secs)

Next, you will assess a second offensive scenario.

Offensive Scenario 02, Task 1:

Task: “Your first task of the second offensive scenario is to draw the route you would have the attacking squad take to building H4.”

Condition: “Given a classroom environment, data collection work sheets, scratch papers, maps of the scenario area, writing utensils, and a ground combat simulation running a **computer/soldier** generated scenario consisting of a dismounted infantry squad conducting operations in an urban environment, assess the behaviors presented on the screen. During the scenario, the model will be paused to allow you time to record your observations. Reminder the higher the value used (1-7 scale) the better the performance.

Standards: “Assess the behaviors utilizing the worksheets provided, your knowledge of urban operations, FM 90-10-1: Combined Arms Operations in Urban Terrain, ARTEP 7-8-MTP: Mission Training Plan for the Infantry Rifle Platoon and Squad, FM 101-5-1: Operational Terms and Graphics, and the assessment procedures you have been taught just prior to this exercise.”

“You have three minutes. Ready,... Begin.”

Offensive Scenario 02, Task 1 Complete

Record Time (to secs)

Offensive Scenario 02, Task 2:

Task: “Your second task is to assess the individual and squad behaviors as they ‘Conduct Tactical Movement in a Built-up Area’ (Infantry Squad) (07-3-1279). Conditions and standards are unchanged.”

“You have three minutes. Ready,... Begin.”

Offensive Scenario 02, Task 2 Complete

Record Time (to secs)

During scenario assessment, I will fast forward through the scenario one time. I will then play the scenario a second time through at near real time speeds. In order to facilitate your ability to record information on the Data Collection Sheets, I will occasionally pause the scenario after the execution of selected tasks and ask you to record your observations.

Record Time (to secs)

Offensive Scenario 02, Task 3.

Task: “Your third task is to assess the individual and squad behaviors as they ‘React to Snipers’ (Infantry Squad) (07-3-1406). Conditions and standards are unchanged.”

“You have three minutes. Ready,... Begin.”

Offensive Scenario 02, Task 3 Complete

Record Time (to secs)

Offensive Scenario 02, Task 4.

Task: “Your forth task is to assess the individual and squad behaviors as they ‘Conduct Tactical Movement in a Built-up Area’ (Infantry Squad) (07-3-1279). Conditions and standards are unchanged.”

“You have three minutes. Ready,... Begin.”

Offensive Scenario 02, Task 4 Complete

Record Time (to secs)

Offensive Scenario 2 Task 5.

Task: “Your fifth task is to assess the overall individual and squad behaviors for this scenario utilizing the last three tasks and other observations you have made. Conditions and standards are unchanged.”

“You have three minutes. Ready,... Begin.”

Offensive Scenario 02, Task 5 Complete

Record Time (to secs)

Overall Assessment, Task 1.

Task: “Your final task is to assess the overall individual and squad behaviors for this study utilizing the last three scenarios and other observations you have made. Conditions and standards are unchanged.”

“You have three minutes. Ready,... Begin.”

Overall Assessment, Task 1 Complete

FINISH

Stop timer

Record Time (to secs)

“Congratulations on completing the final task and scenario for this experiment.”

9. DEBRIEF

You have just completed a data collection phase for a study designed to identify methods for potentially assisting in the development and validation of computer models designed to replicate human performance.

We ask you to complete the debriefing questionnaire to assist us in the design and conduct of future studies. We also ask that you refrain from discussing this study with anyone other than the personnel collecting data for the study. This will help to ensure that potential participants will not be biased by your experiences and thus corrupt our data.

On behave of the MOVES Institute, the Naval Postgraduate School, and the Department of Defense; we thank you for your time and comments.

APPENDIX M. EXPERIMENT BRIEFING SLIDES

For this research's experiments, participants viewed two Power Point presentations. The first set of slides was used on day one of the experiment for the in-processing, familiarization, and training phases of the study. The second series of slides was used on the second day for the refresher, data collection, and debriefing phases of the study. Each study uses the same basic slides with modifications for the assessment scales. This appendix provides the two different representative slide sets used for the two experiments.

The following are the two sets of slides utilizing the 7-Point Likert Scale used for the first study. The second study used four sets of slides. One series of presentations (two sets) is used for the participants utilizing the 5-Point Likert Scale. The second series of slides (two sets) is used for the participants utilizing the Go/No-Go scale. As with the first experiment, one set of slides for each group of participants, is used on the first day of the experiment and the second set of slides is used on day two of the experiment.

1. SLIDE SET FOR DAY #1, 7-POINT LIKERT SCALE

The following are the 32 slides used on the first day of the study.



Validating Human Behavioral Models for Combat Simulations Using Techniques for the Evaluation of Human Performance

**Study 01, Day 01
18-24 September 2003
Fort Benning, GA**

MAJ Simon R. Goerge, USA
srgoerge@nps.navy.mil
(831) 656 - 3733

1

Outline

- Motivation & Goal
- Study Phases
- In-Processing
- Assessment Procedures
- Operations Overview
- Practical Exercise
- Review
- Assessment
- Debriefing

2

Motivation & Goal

- Motivation
 - Validation for behavioral representation models is not well defined, nor is the current process extensible to meet requirements for validating the varied and complex human behavior representation and cognitive models in use or under development for Department of Defense (DoD) simulations
- Goal
 - To outline some of the issues with validating cognitive models for use by the DoD Modeling and Simulation (M&S) community and to propose potential means for mitigating these issues

3

Study Phases

- In-Processing (*Day #1, 15 Minutes*)
 - Participant Demographics Questionnaire
 - Participant Consent Form
 - Minimal Risk Consent Form
 - Privacy Act Statement
 - NEO Five Factor Personality Test
- Training Phase (*Day #1, 25 Minutes*)
 - Assessment Procedures
 - Model Familiarization
 - Operations Overview
- Participant Practice (*Day #1, 20 Minutes*)
- Review Phase (*Day #2, 10 Minutes*)
- Assessment Phase (*Day #2, 45 Minutes*)
- Debriefing (*Day #2, 5 Minutes*)

4

In-Processing

- Participant Demographics Questionnaire
- Participant Consent Form
- Minimal Risk Consent Form
- Privacy Act Statement
- NEO Five Factor Personality Test

5

Assessment Procedures

- Steps
 - Assess performance of task
 - ♦ Pre-Run Map Sketch(s)
 - ♦ Task Assessment Checklist(s)
 - Assess performance of scenario
 - ♦ Scenario Assessment Checklist
 - Assess overall performance
 - ♦ Study Assessment Checklist

6

Assessment Procedures

• Example Pre-Run Map Sketch



Weapons Symbols

Mines	
Antipersonnel (AP)	
Antitank (AT)	

Wire Obstacles	
Unspecified	XXXXXXXXXX
Triple Stand Concertina	

Weapons		
M16/M24		
M203		
M249		
AT4		
60mm Mortar		

Assessment Procedures

• Assessment Forms

- Modified check lists
 - ARTEP 7-8-MTP: Mission Training Plan for the Infantry Rifle Platoon and Squad
- Likert seven factor evaluation scale with not applicable (NA) and comments section
- Sub Task Summary Table and Task Summary
- Task Summary, Task Summary Table, and Scenario Summary
- Scenario Summary and Scenario Summary Table
- Study Overall Assessment

9

Assessment Procedures

• Likert seven factor evaluation scale with not applicable (NA)

- Low score, means low or sub pare performance
- The higher the score, the better the performance

ASSESSMENT SCALE: Use the following scale to assess performance of the squad as it performs this task.

- 1 - **Strongly agree** the task, step, or performance measure was *improperly* performed
 - 2 - **Agree** the task, step, or performance measure was *improperly* performed
 - 3 - **Not sure** but tend to agree the task, step, or performance measure was *improperly* performed
 - 4 - **Undecided**
 - 5 - **Not sure** but tend to agree the task, step, or performance measure was *properly* performed
 - 6 - **Agree** the task, step, or performance measure was *properly* performed
 - 7 - **Strongly agree** the task, step, or performance measure was *properly* performed
- NA - Not applicable or no means of determining

10

Assessment Procedures

• Example of Likert seven factor evaluation scale with not applicable (NA) and comments section

OFFENSIVE SCENARIO, TASK: Conduct Tactical Movement in a Built-up Area (Infantry Squad) (07-3-1279)	
TASK(S), STEP(S) and PERFORMANCE MEASURE(S)	ASSESSMENT
2. Squad moves only after defenders' fires have been suppressed or obscured, if applicable. <i>Comments:</i>	1-2-3-4-6-7 NA
3. Squad moves at night or during other periods of reduced visibility using night vision devices (NVDs). <i>Comments:</i> Due to slow movement to the town, forces were required to attack the town in daylight.	1-2-3-4-5-6-7 NA
4. Squad moves using concealment of smoke provided by supporting vehicles or assets. <i>Comments:</i> Did not use smoke, but I am not sure any was available form supporting assets.	1-2-3-4-5-6-7 NA

11

Assessment Procedures

• Example Comments

- Cannot tell if they could see enemy personnel
- Squad leader used lead team to lay down a base of fire, fixing the sniper so the second team could maneuver to close with and kill the sniper
- Squad leader never spoke with his higher headquarters.

12

Assessment Procedures

- Sub Task Summary Table and Task Summary

TASK PERFORMANCE SUMMARY BLOCK								
ASSESSMENT CATEGORY	1	2	3	4	5	6	7	NA
Total tasks, steps, and performance measures evaluated at this level.	2	4	4	3	4	3	1	5
TASK: Conduct Tactical Movement in a Built-up Area (Infantry Squad) (07-3-1279)								
Squad performance this task is standard. Comments: Although the many sub tasks were not performed because they squad never reached the OBJ, I feel they would have been performed to an acceptable standards.								
ASSESSMENT								
1-2-3-4-6-7 NA								

13

Assessment Procedures

- Task Summary, Task Summary Table, and Scenario Summary

TASKS, STEPS, AND PERFORMANCE MEASURES								
1. Conduct Tactical Movement in a Built-up Area (Infantry Squad) (07-3-1279)	1-2-3-4-6-7	NA	0	0	0	0	0	0
2. Squad Leader fails to control movement of team to ensure they provide supporting fire.	1-2-3-4-6-7	NA	0	0	0	0	0	0
3. Squad Leader fails to bring in-team around to use for supporting fire or call for assets.	1-2-3-4-6-7	NA	0	0	0	0	0	0
4. Conduct Tactical Movement in a Built-up Area (Infantry Squad) (07-3-1279)	1-2-3-4-6-7	NA	0	0	0	0	0	0
5. Squad Leader fails to control movement of team to ensure they provide supporting fire.	1-2-3-4-6-7	NA	0	0	0	0	0	0
6. Squad Leader fails to bring in-team around to use for supporting fire or call for assets.	1-2-3-4-6-7	NA	0	0	0	0	0	0
7. Squad Leader fails to control movement of team to ensure they provide supporting fire.	1-2-3-4-6-7	NA	0	0	0	0	0	0
8. Squad Leader fails to bring in-team around to use for supporting fire or call for assets.	1-2-3-4-6-7	NA	0	0	0	0	0	0
9. Squad Leader fails to control movement of team to ensure they provide supporting fire.	1-2-3-4-6-7	NA	0	0	0	0	0	0
10. Squad Leader fails to bring in-team around to use for supporting fire or call for assets.	1-2-3-4-6-7	NA	0	0	0	0	0	0
TASK PERFORMANCE SUMMARY BLOCK								
ASSESSMENT CATEGORY	1	2	3	4	5	6	7	NA
Total tasks, steps, and performance measures evaluated at this level.	0	0	1	1	2	0	0	1
OFFENSIVE SCENARIO #1								
Squad performance this task is standard. Comments: Individual and team movement was good, but the squad leader needs to control the team and focus combat power quicker.								
ASSESSMENT								
1-2-3-4-6-7 NA								

14

Assessment Procedures

- Scenario Summary and Scenario Summary Table

TASKS, STEPS, AND PERFORMANCE MEASURES								
1. Defensive Scenario #1	1-2-3-4-6-7	NA	0	0	0	0	0	0
2. Defensive Scenario #1	1-2-3-4-6-7	NA	0	0	0	0	0	0
3. Defensive Scenario #2	1-2-3-4-6-7	NA	0	0	0	0	0	0
4. Defensive Scenario #2	1-2-3-4-6-7	NA	0	0	0	0	0	0
NOTE: * Indicates a leader task.								
TASK PERFORMANCE SUMMARY BLOCK								
ASSESSMENT CATEGORY	1	2	3	4	5	6	7	NA
Total tasks, steps, and performance measures evaluated at this level.	0	0	1	0	2	0	0	1

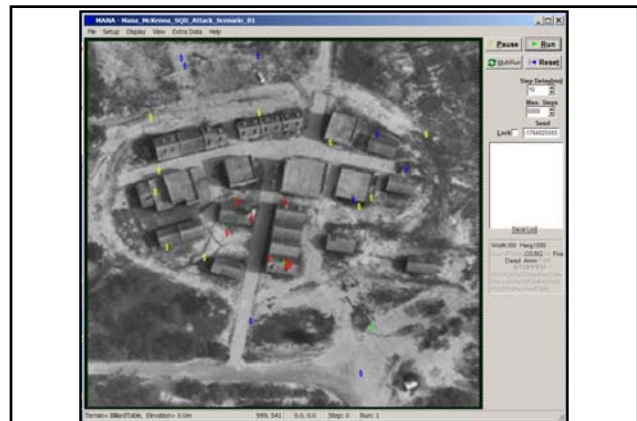
Assessment Procedures

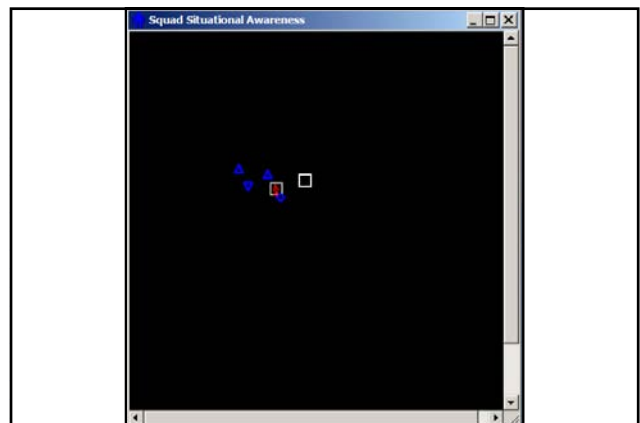
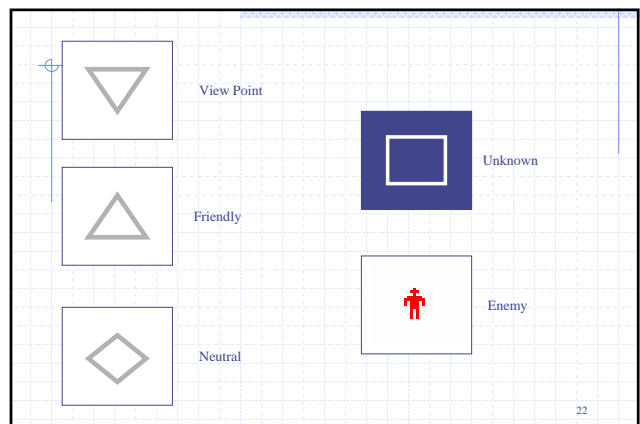
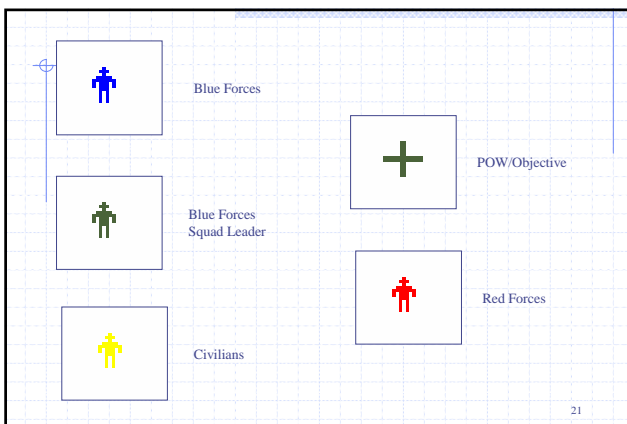
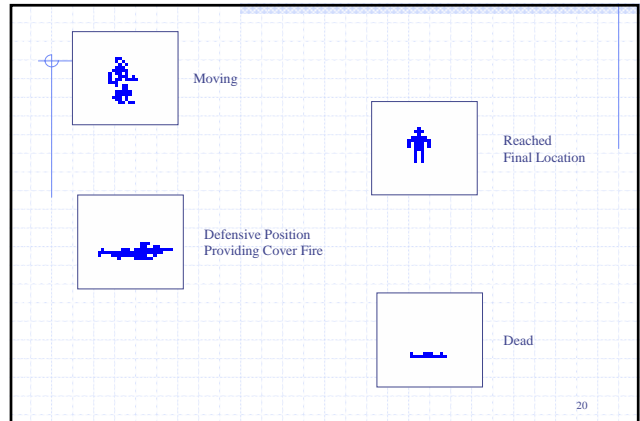
- Study Overall Assessment
 - Performance
 - Future Performance

General Performance								
At the individual squad member level, the squad model received the comment(s) in a manner appropriate for US forces operating in urban terrain.	1-2-3-4-6-7	NA	0	0	0	0	0	0
Individual movement and weapons proficiency were outstanding. Squad members sought cover and concealment without supervision and followed the orders from their leaders.	1-2-3-4-6-7	NA	0	0	0	0	0	0
At the squad level, the squad model received the comment(s) in a manner appropriate for US forces operating in urban terrain.	1-2-3-4-6-7	NA	0	0	0	0	0	0
By the end of the evaluation, the squad leader could control his team and focus his combat power at the decisive point.	1-2-3-4-6-7	NA	0	0	0	0	0	0
General Performance								
I am confident the squad model can execute individual tasks in a manner appropriate for US forces operating in urban terrain.	1-2-3-4-6-7	NA	0	0	0	0	0	0
Individual movement techniques may need modification for more open terrain. I have confidence squad members will seek cover and concealment with minimal supervision and will follow orders from their leaders.	1-2-3-4-6-7	NA	0	0	0	0	0	0
I am confident the squad model can execute squad level tasks in a manner appropriate for US forces operating in urban terrain.	1-2-3-4-6-7	NA	0	0	0	0	0	0
The squad leader needs more practice with developing situational awareness of different environments and adapting his use of his team based on the situation.	1-2-3-4-6-7	NA	0	0	0	0	0	0

Model Familiarization

- MANA Interface
- Terrain
- Symbols
 - Posture
 - Forces
 - Situational Awareness
- Situational Awareness Map





Operations Overview

- Urban Operation
- ROE
- “Cross Roads for the Night”

25

Rules of Engagement (ROE)

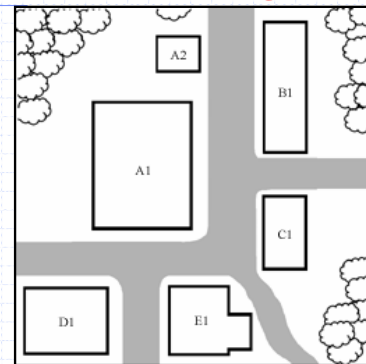
- Take all steps necessary and appropriate for your unit's protection.
- The minimum necessary force will be used to control the situation. Rifles will be placed in single shot mode to reduce fratricide, civilian casualties, and excess use of ammunition.
- To reduce friendly casualties and damage to the buildings, we will use concussion grenades within the boundaries of the village. If you deploy a dummy Flash Bang, there will be a 2-second delay before detonation. Concussion grenades will incapacitate personnel for approximately 5 seconds, thus swift movement to secure enemy personnel is essential.
- Follow standard MOUT Tactics, Techniques, and Procedures (TTPs).
- Take measures to minimize risk to civilians, without endangering the unit.
- Fire is returned directly to its source, not sprayed into a general area (use single shoot selection for rifles).
- Firing ceases when the threat is over.
- Anyone trying to surrender is allowed to do so.
- Civilians and property are treated with respect.
- WP can be used vicinity the town to aid in isolating the objectives. The requests for indirect fire within the town must be authorized by the battalion commander.
- No use of artillery inside the town.

Participant Practice

- Task:
 - Conduct a Strongpoint Defense of a Building (Infantry Squad)
- Condition:
 - Given a classroom environment, projected display of a scenario displayed in MANA, sketches, record sheet, scratch papers, pen, and a ground combat simulation running a scenario consisting of a dismounted infantry squad conducting operations in an urban environment, assess the behaviors presented on the screen. During the scenario, the model will be paused to allow you time to record your observations.
- Standards: Assess the behaviors utilizing:
 - Modified MTP worksheets
 - Your knowledge of urban operations
 - FM 90-10-1: Combined Arms Operations in Urban Terrain
 - ARTEP 7-8-MTP: Mission Training Plan for the Infantry Rifle Platoon and Squad
 - Assessment procedures you have been taught for this study

27

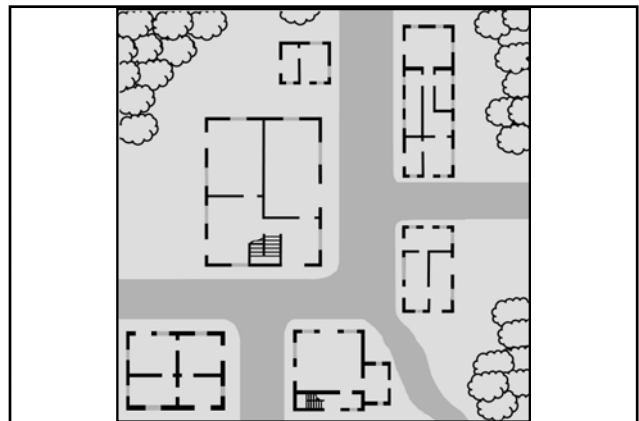
“Cross Roads for the Night” Sketch


















28

“Cross Roads for the Night” Terrain

- Road
- Unrestricted (Open Terrain)
- Unrestricted (Open Terrain)
- Restricted
- Severely Restricted
- Wall



Weapons Symbols	
Mines	
Antipersonnel (AP)	
Antitank (AT)	
Wire Obstacles	
Unspecified	XXXXXXXXXX
Triple Stand Concertina	
Weapons	
M16/M24 (6)	 
M203 (2)	 
M249 (2)	 
AT4 (4 located w/M16)	 
60mm Mortar (0)	 

Validating Human Behavioral Models for Combat Simulations Using Techniques for the Evaluation of Human Performance



Study 01, Day 01
18-24 September 2003
Fort Benning, GA

MAJ Simon R. Goerger, USA
srgoerge@nps.navy.mil
(831) 656 - 3733

32

2. SLIDE SET FOR DAY #2, 7-POINT LIKERT SCALE

The following are the 30 slides used on the refresher and day of the study.






Validating Human Behavioral Models for Combat Simulations Using Techniques for the Evaluation of Human Performance

Study 01, Day 02
18-24 September 2003
Fort Benning, GA

MAJ Simon R. Goerge, USA
 srgoerge@nps.navy.mil
 (831) 656 - 3733



1

Outline

- Motivation & Goal
- Study Phases
- In-Processing
- Assessment Procedures
- Operations Overview
- Practical Exercise
- Review
- Assessment
- Debriefing



2

Participant Phases

- In-Processing (*Day #1, 15 Minutes*)
 - Participant Demographics Questionnaire
 - Participant Consent Form
 - Minimal Risk Consent Form
 - Privacy Act Statement
 - NEO Five Factor Personality Test
- Training Phase (*Day #1, 25 Minutes*)
 - Assessment Procedures
 - Model Familiarization
 - Operations Overview
- Participant Practice (*Day #1, 20 Minutes*)
- Review Phase (*Day #2, 10 Minutes*)
- Assessment Phase (*Day #2, 45 Minutes*)
- Debriefing (*Day #2, 5 Minutes*)



3

Assessment Procedures


- Steps
 - Assess performance of task
 - ♦ Pre-Run Map Sketch(s)
 - ♦ Task Assessment Checklist(s)
 - Assess performance of scenario
 - ♦ Scenario Assessment Checklist
 - Assess overall performance
 - ♦ Study Assessment Checklist

4






Assessment Procedures

•Example Pre-Run Map Sketch



Weapons		
M16/M24	↑	↑
M203	⊕	⊕
M249	↑	↑
AT4	↑	↑
60mm Mortar	⊕	⊕

Assessment Procedures

- Likert seven factor evaluation scale with not applicable (NA)
 - Low score, means low or sub pare performance
 - The higher the score, the better the performance

ASSESSMENT SCALE: Use the following scale to assess performance of the squad as it performs this task.

1 – Strongly agree the task, step, or performance measure was *improperly* performed
 2 – Agree the task, step, or performance measure was *improperly* performed
 3 – Not sure but tend to agree the task, step, or performance measure was *improperly* performed
 4 – Undecided
 5 – Not sure but tend to agree the task, step, or performance measure was *properly* performed
 6 – Agree the task, step, or performance measure was *properly* performed
 7 – Strongly agree the task, step, or performance measure was *properly* performed
 NA - Not applicable or no means of determining

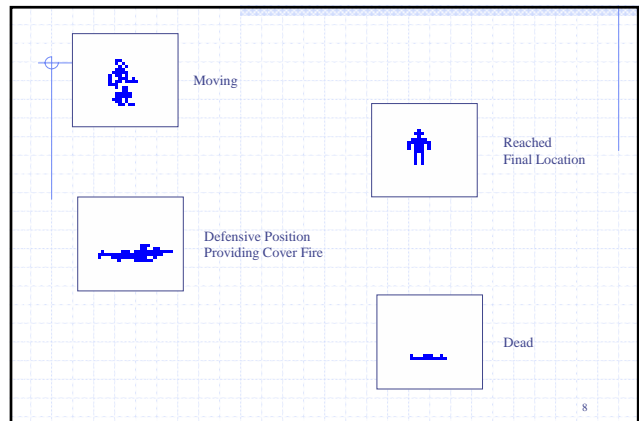
6

Assessment Procedures

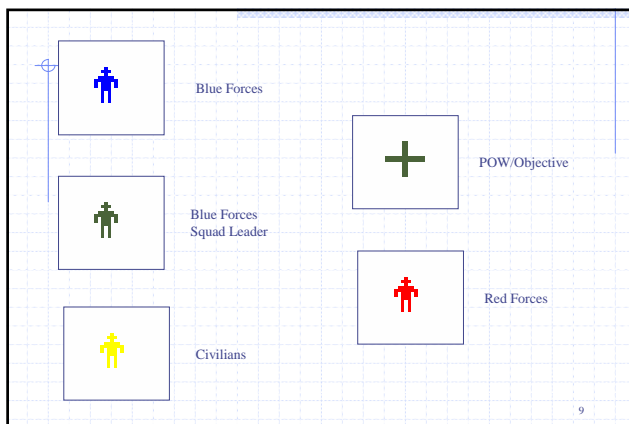
- Example Comments

- Cannot tell if they could see enemy personnel
- Squad leader used lead team to lay down a base of fire, fixing the sniper so the second team could maneuver to close with and kill the sniper
- Squad leader never spoke with his higher headquarters.

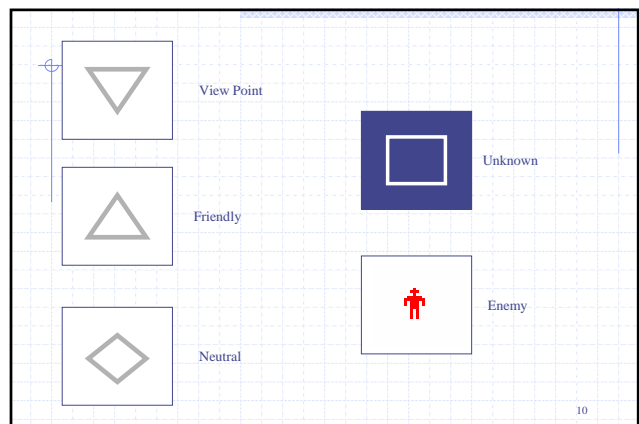
7



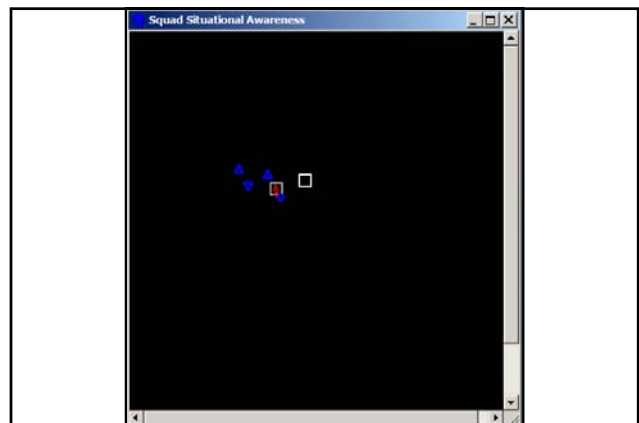
8



9



10



Assessment

- Scenario #1 – Offensive operations
 - Conduct Tactical Movement in a Built-up Area
 - React to Snipers
 - Conduct Tactical Movement in a Built-up Area
 - Scenario Assessment
- Scenario #2 – Defensive operations
 - Conduct a Strongpoint Defense of a Building
 - Scenario Assessment
- Scenario #3 – Offensive operations
 - Conduct Tactical Movement in a Built-up Area
 - React to Snipers
 - Conduct Tactical Movement in a Built-up Area
 - Scenario Assessment
- Overall Performance Assessment

13

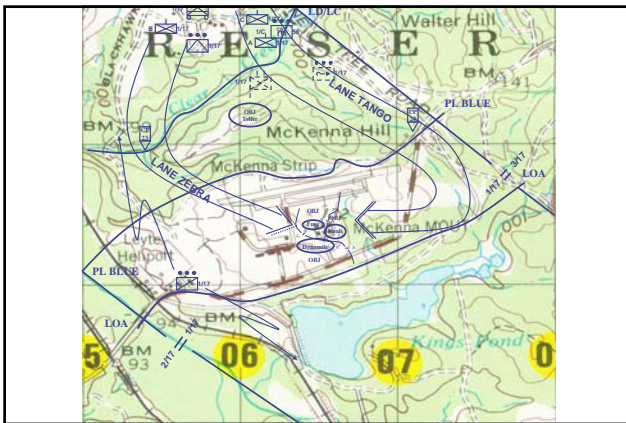
Operations Overview

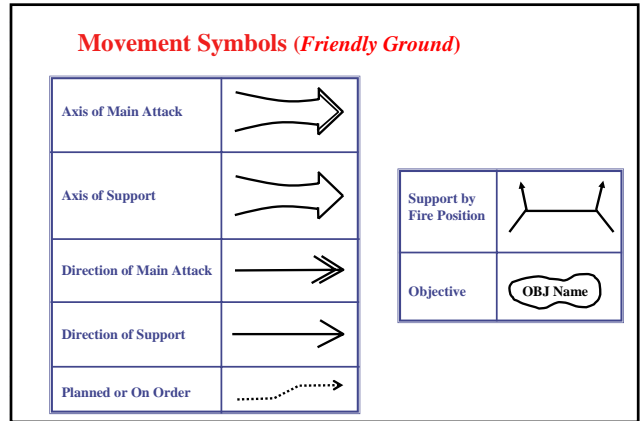
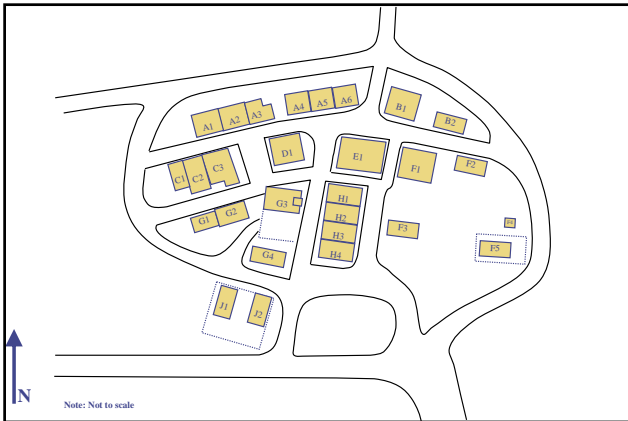
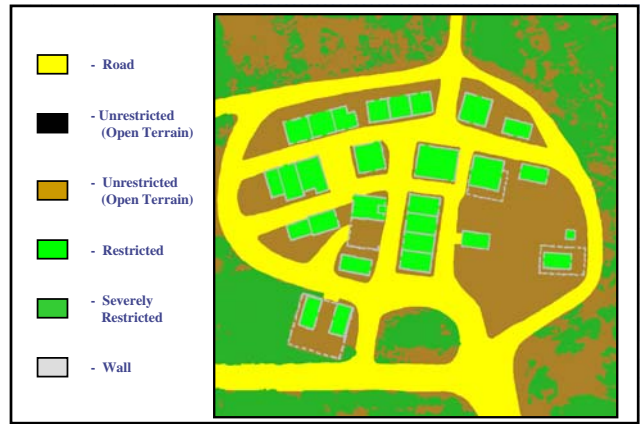
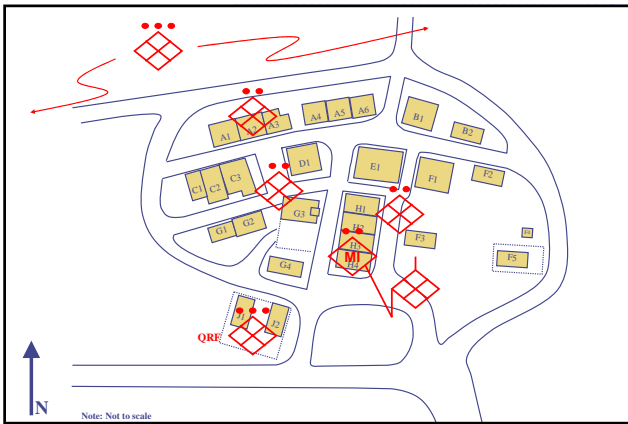
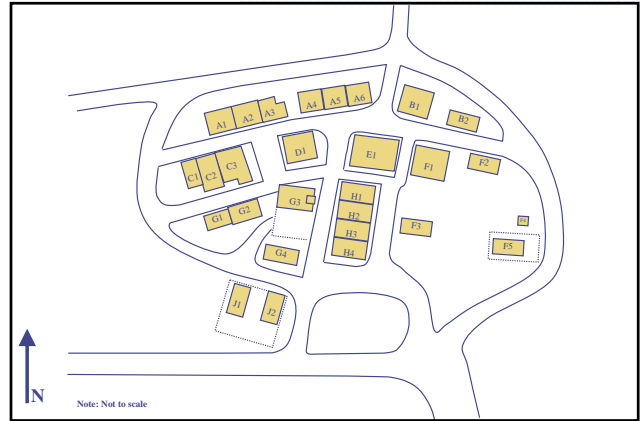
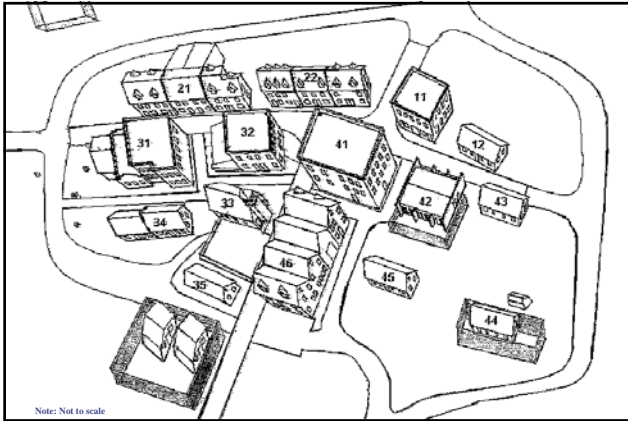
- Urban Operation
- ROE
- McKenna, GA

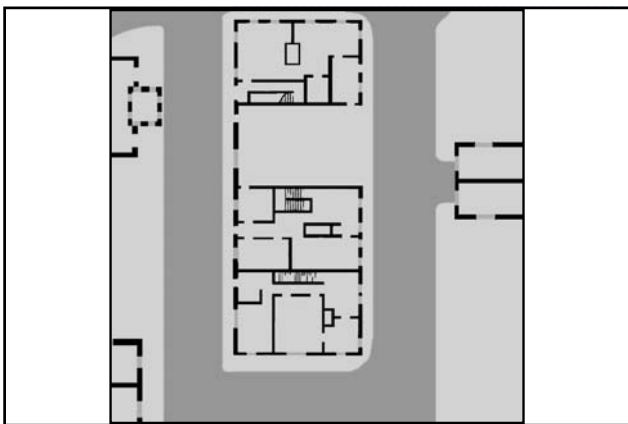
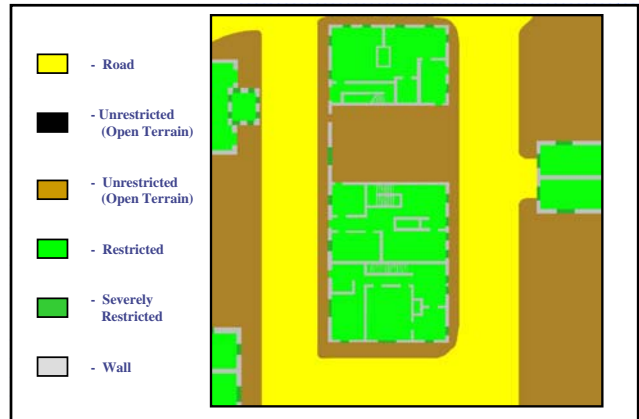
14

Rules of Engagement (ROE)

- Take all steps necessary and appropriate for your unit's protection.
- The minimum necessary force will be used to control the situation. Rifles will be placed in single shot mode to reduce fratricide, civilian casualties, and excess use of ammunition.
- To reduce friendly casualties and damage to the buildings, we will use concussion grenades within the boundaries of the village. If you deploy a dummy Flash Bang, there will be a 2-second delay before detonation. Concussion grenades will incapacitate personnel for approximately 5 seconds, thus swift movement to secure enemy personnel is essential.
- Follow standard MOUT Tactics, Techniques, and Procedures (TTPs).
- Take measures to minimize risk to civilians, without endangering the unit.
- Fire is returned directly to its source, not sprayed into a general area (use single shoot selection for rifles).
- Firing ceases when the threat is over.
- Anyone trying to surrender is allowed to do so.
- Civilians and property are treated with respect.
- WP can be used vicinity the town to aid in isolating the objectives. The requests for indirect fire within the town must be authorized by the battalion commander.
- No use of artillery inside the town.







Weapons Symbols



Mines	
Antipersonnel (AP)	
Antitank (AT)	

Wire Obstacles	
Unspecified	XXXXXXXXXX
Triple Stand Concertina	

Weapons		
M16/M24 (6)		
M203 (2)		
M249 (2)		
AT4 (4 located w/M16)		
60mm Mortar (0)		

Debriefing

- Debriefing Handout
- NEO Five Factor Inventory Results
- America's Army CDs

Validating Human Behavioral Models for Combat Simulations Using Techniques for the Evaluation of Human Performance

Study 01, Day 02
18-24 September 2003
Fort Benning, GA

MAJ Simon R. Goerger, USA
 srgoerge@nps.navy.mil
 (831) 656 - 3733

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX N. SUPPORTING FIGURES AND TABLES FOR DATA ANALYSIS

The following are figures and tables from the analysis of data for the experiment conducted in support of the research on subject matter expert (SME) bias in the assessment of cognitive models that do not appear in the main body of this dissertation. They display the results from various analysis techniques using The Statistical Discovery Software, JMP® from SAS Institute Inc. Some figures are developed using the Excel® spreadsheet functionality.

1. PARTICIPANT DEMOGRAPHICS

Table N.1. Participant Demographics: Education & Service Data¹⁸³

Statistic	Assessment Scale			
	All Scales	7-Point	Go/No-Go	5-Point
Total Number of Participants using Assessment Scale	182	80	50	52
Highest Level of Education				
Number of Bachelors	173	75	49	49
Number of Masters	6	3	1	2
Component				
Number of Active Duty	172	74	47	51
Number of National Guard	8	4	3	1
Service				
Number in Army	178	79	50	49
Number in Marines	4	1	0	3
Rank				
Number of First Lieutenants	8	3	1	4
Number of Captains	174	77	49	48
Branch (Primary)				
Number in Infantry	161	78	41	40
Number in Special Forces	1	0	1	0
Number in Other Combat Arms ¹⁸⁴	3	0	2	1
Number in Other Combat Support ¹⁸⁵	11	0	5	6
Number in Other Combat Service	3	0	1	2

¹⁸³ Data excludes participants who did not respond to the specific question(s) on the Participant Demographics Questionnaire.

¹⁸⁴ The Combat Arms category normally includes infantry; however, infantry officers are categorized separately in this table and combat arms for this table include Armor, Aviation, and Special Forces.

¹⁸⁵ Combat Support branches include Air Defense, Engineers, Field Artillery, Military Police, and Military Intelligence. Many of these personnel are slated to go through the Special Forces Qualification Course after graduating ICCC.

Statistic	Assessment Scale			
	All Scales	7-Point	Go/No-Go	5-Point
Support ¹⁸⁶				
Unit¹⁸⁷				
Light Infantry	39	21	10	8
Mechanized Infantry	58	26	16	16
Air Assault	38	19	10	9
Airborne	51	20	16	15
Special Operations ¹⁸⁸	23	10	5	8
Other ¹⁸⁹	31	12	11	8
Duty Position¹⁹⁰				
Automatic Rifleman	31	14	9	8
Grenadier	26	112	7	7
Fire Team Leader	28	12	9	7
Squad Leader	29	12	12	5
Rifle Platoon Leader	149	73	35	41
Scout Platoon Leader	29	15	7	7
Anti-Tank Platoon Leader	32	17	6	9
Mortar Platoon Leader	27	13	10	4
Rifle Company Executive Officer	90	46	24	20
Rifle Company Commander	2	2	0	0
Operations Staff Officer	66	22	22	22
Other	99	40	29	30

Table N.2. NEO-FFI Raw Score Conversions From [COST 92]

Level	Raw Score Values				
	Neuroticism (N)	Extraversion (E)	Openness (O)	Agreeableness (A)	Conscientiousness (C)
Very Low	0 - 6	0 - 18	0 - 18	0 - 24	0 - 25
Low	7 - 13	19 - 24	19 - 23	25 - 29	25 - 30
Average	16 - 21	25 - 30	24 - 30	30 - 34	31 - 37
High	22 - 29	31 - 36	31 - 36	25 - 40	38 - 43
Very High	30 - 50	37 - 50	37 - 50	41 - 50	44 - 50

¹⁸⁶ Combat Service Support branches include Signal Corps and Transportation Corps officers. These personnel are slated to go through the Special Forces Qualification Course after graduating ICCC.

¹⁸⁷ Participants may have served in more than one type of unit.

¹⁸⁸ Special Operations includes Rangers, Special Forces, and Delta Force units.

¹⁸⁹ Other units include armor, cavalry, Striker Brigade, artillery, engineer, general support, etc.

¹⁹⁰ Participants may have served in more than one duty position.

Table N.3. Participant NEO-FFI Raw Score Statistics

Statistic	Assessment Scale				US - Men ¹⁹¹
	Total	7-Point	Go/No-Go	5-Point	
Total Number (N) of Participants using Assessment Scale	182	80	50	52	
Neuroticism (N) Raw-score - N	179	80	50	49	
Min	0	1	0	1	
Max	30	29	30	23	
Mean	11.93	12.54	12.66	10.18	17.60
Std Dev	6.54 ¹⁴	6.52	7.22	5.60	7.46
Extraversion (E) Raw-score - N	179	80	50	49	
Min	18	21	18	23	
Max	47	45	47	40	
Mean	32.10	32.03	31.86	32.47	27.22
Std Dev	4.83	4.85	5.44	4.15	5.85
Openness (O) Raw-score - N	179	80	50	49	
Min	13	13	17	16	
Max	40	40	40	40	
Mean	27.27	26.69	27.48	28.00	27.09
Std Dev	5.76	5.66	5.65	6.05	5.82
Agreeableness (A) Raw-score - N	179	80	50	49	
Min	12	18	12	14	
Max	43	43	41	42	
Mean	30.13	30.06	29.74	30.63	31.93
Std Dev	5.86	5.90	5.63	6.13	5.03
Conscientiousness (C) Raw-score - N	179	80	50	49	
Min	18	18	21	23	
Max	48	47	46	48	
Mean	35.28	34.70	35.14	36.37	34.10
Std Dev	5.79 ¹⁹²	5.65	5.94	5.83	5.95

¹⁹¹ The numbers for US – Men are from the *NEO PI-R Professional Manual* [COST 92].

¹⁹² Participant responses are not normal thus, the standard deviation is not a viable value for this data. It is shown here as an approximation for comparison with the US – Men responses.

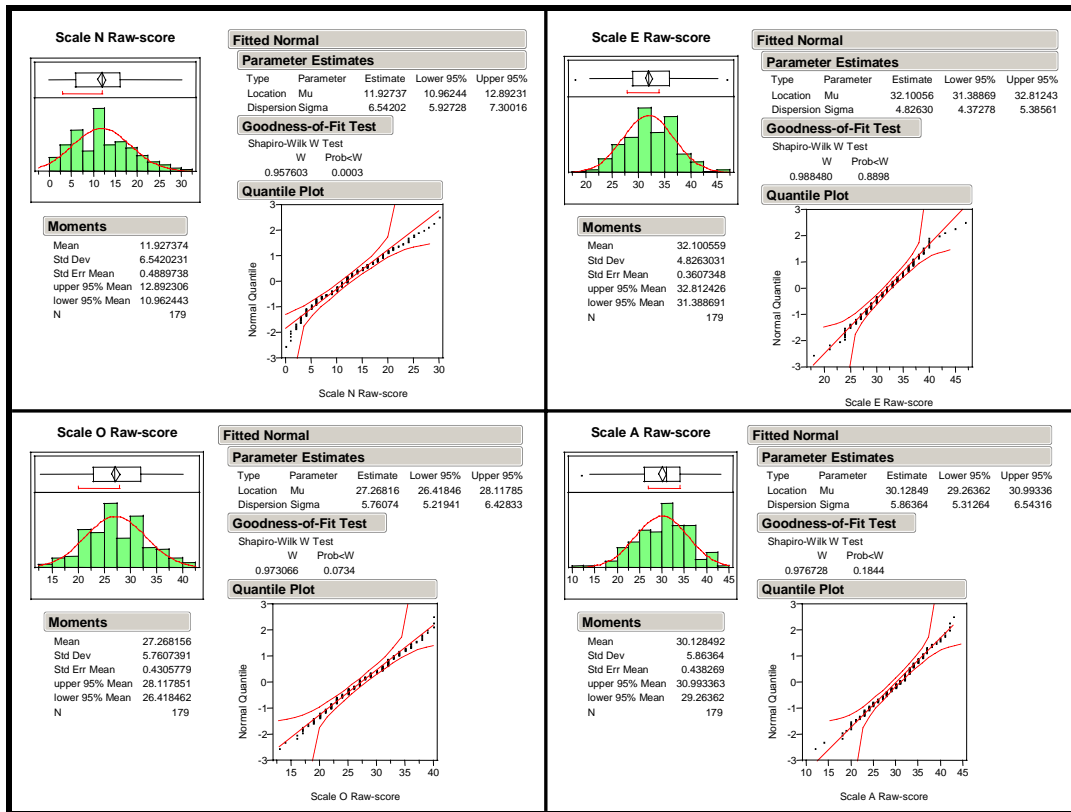


Figure N.1. Participant NEO-FFI Raw Score Fitted Normal Quad Chart

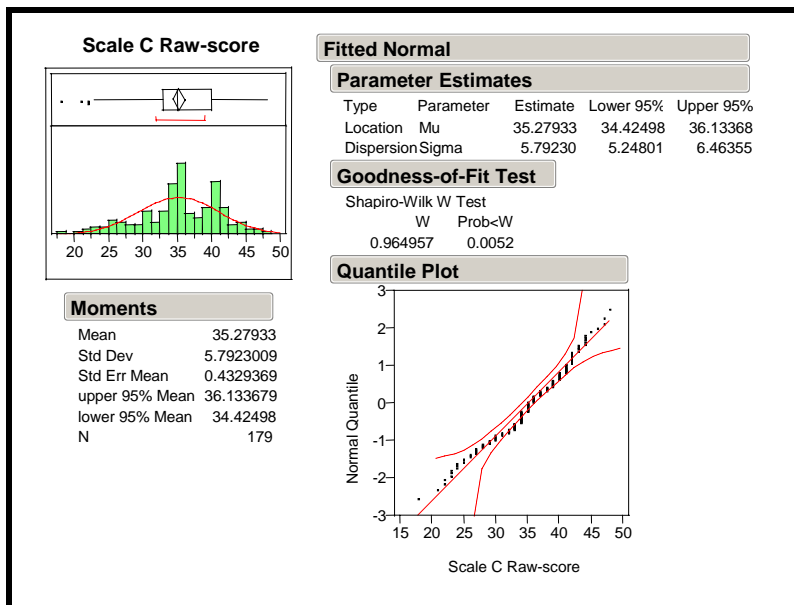


Figure N.2. Participant NEO-FFI Raw Score Fitted Normal Chart - Conscientiousness

2. CONSISTENCY

Table N.4. Likelihood Ratio Tests: Subtask-to-Task Effect - Consistency

Source	Number of Parameters	DF	L-R ChiSquare	Prob>ChiSq
Scale	2	2	125.785802	0.0000
Simulation Belief	1	1	4.21569351	0.0401
Scale * Simulation Belief	2	2	0.46924746	0.7909

Table N.5. Consistency Means of Normalized Values: Subtask-to-Task

Level	Number	Mean
0-1	269	-0.05576
0-2	174	-0.01437
0-3	166	-0.08916
1-1	277	-0.02372
1-2	163	-0.05215
1-3	165	-0.05455

Table N.6. Likelihood Ratio Tests, Task-to-Scenario Effect - Consistency

Source	Number of Parameters	DF	L-R ChiSquare	Prob>ChiSq
Scale	2	2	45.7446681	0.0000
Simulation Belief	1	1	0.02566709	0.8727
Scale * Simulation Belief	2	2	0.9199324	0.6313

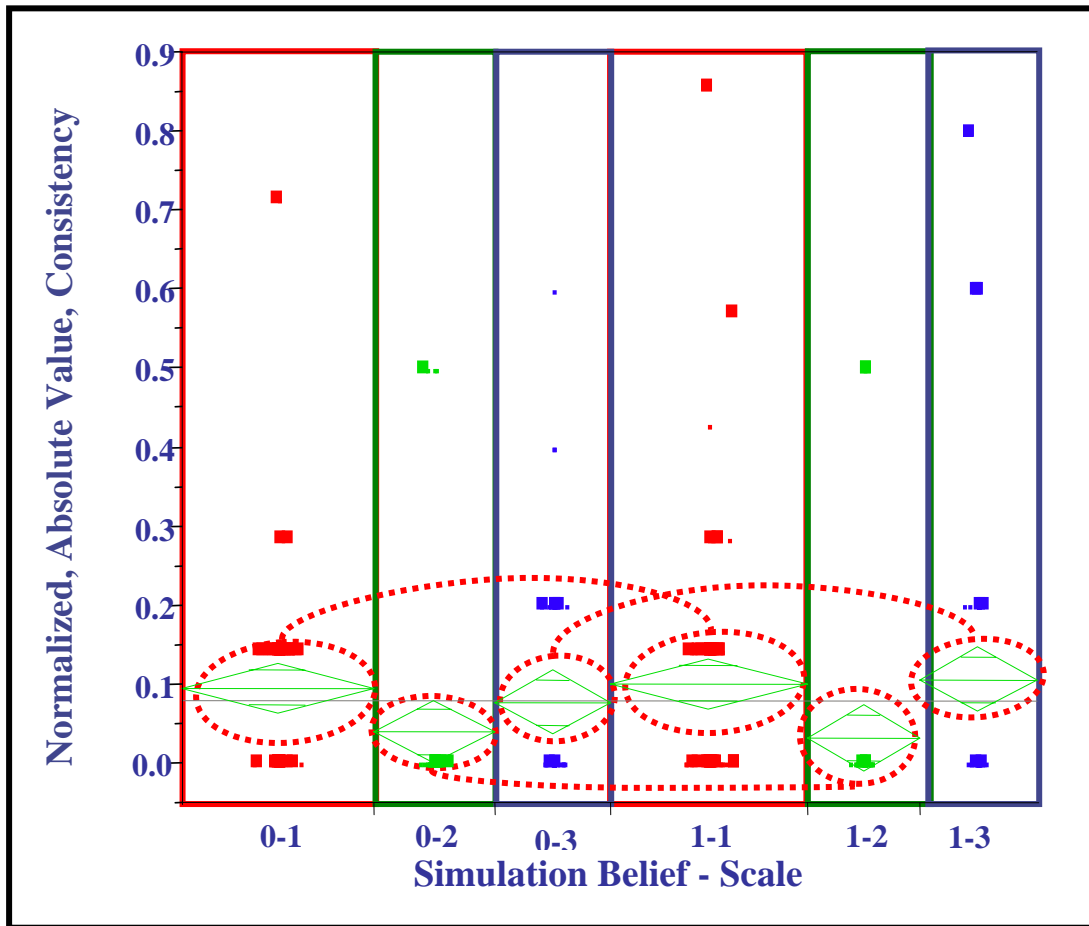


Figure N.3. Intra-SME Task-to-Scenario Consistency Scores

Table N.7. Consistency Means of Normalized Absolute Values: Task-to-Scenario

Level	Number	Mean (Population)
0-1	77	0.096475
0-2	49	0.040816
0-3	46	0.078261
1-1	79	0.101266
1-2	45	0.033333
1-3	47	0.106383

Table N.8. Consistency Means of Normalized Values: Task-to-Scenario

Level	Number	Mean
0-1	77	-0.02226
0-2	49	0.02041
0-3	46	-0.03478
1-1	79	0.01808
1-2	45	-0.01111
1-3	47	-0.03830

Table N.9. Likelihood Ratio Tests, Scenario-to-Overall Effect - Consistency

Source	Number of Parameters	DF	L-R ChiSquare	Prob>ChiSq
Scale	2	2	60.7713191	0.0000
Simulation Belief	1	1	3.56905986	0.0589
Scale * Simulation Belief	2	2	0.30744082	0.8575

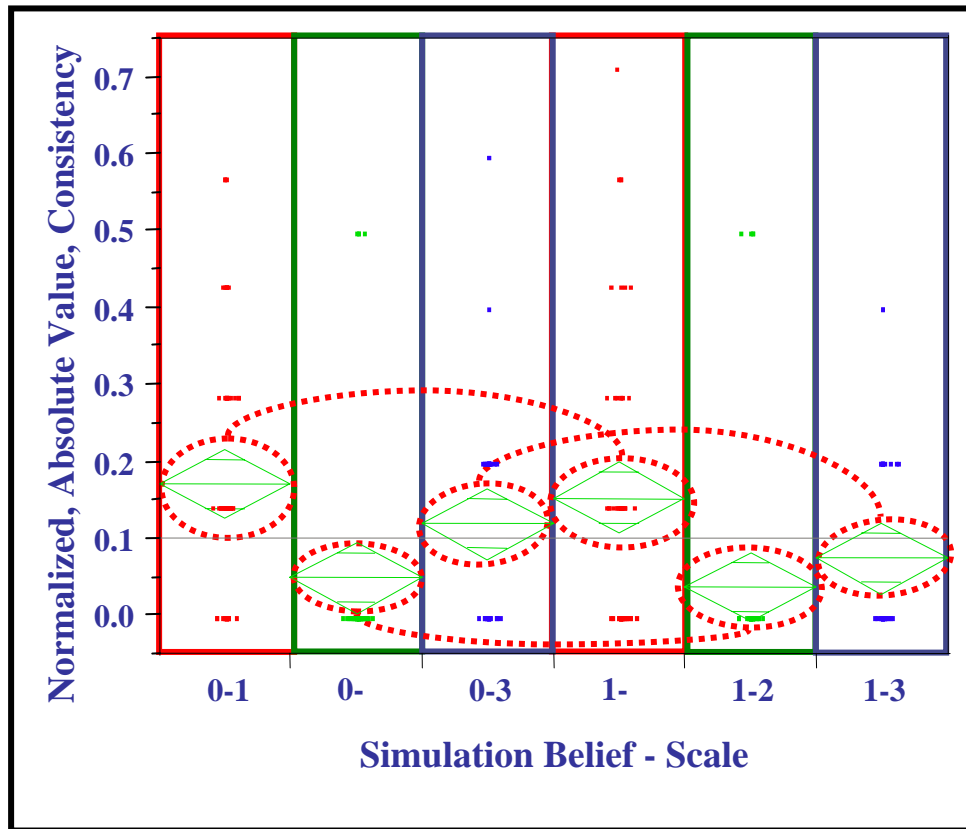


Figure N.4. Intra-SME Scenario-to-Overall Consistency Scores

3. ACCURACY

Table N.10. Ordinal Logistical Fit for Normalized Accuracy Scores

Level	Prob>ChiSq			
	Whole Model Test	Effect Likelihood Ratio Test		
		Scale	Simulation Belief	Scale cross Simulation Belief
Subtask	0.0001	0.0000	0.0116	0.0000
Task	0.0001	0.0000	0.0015	0.1788
Scenario	0.0001	0.0000	0.0038	0.9201
Overall	0.0001	0.0000	0.1216	0.0076

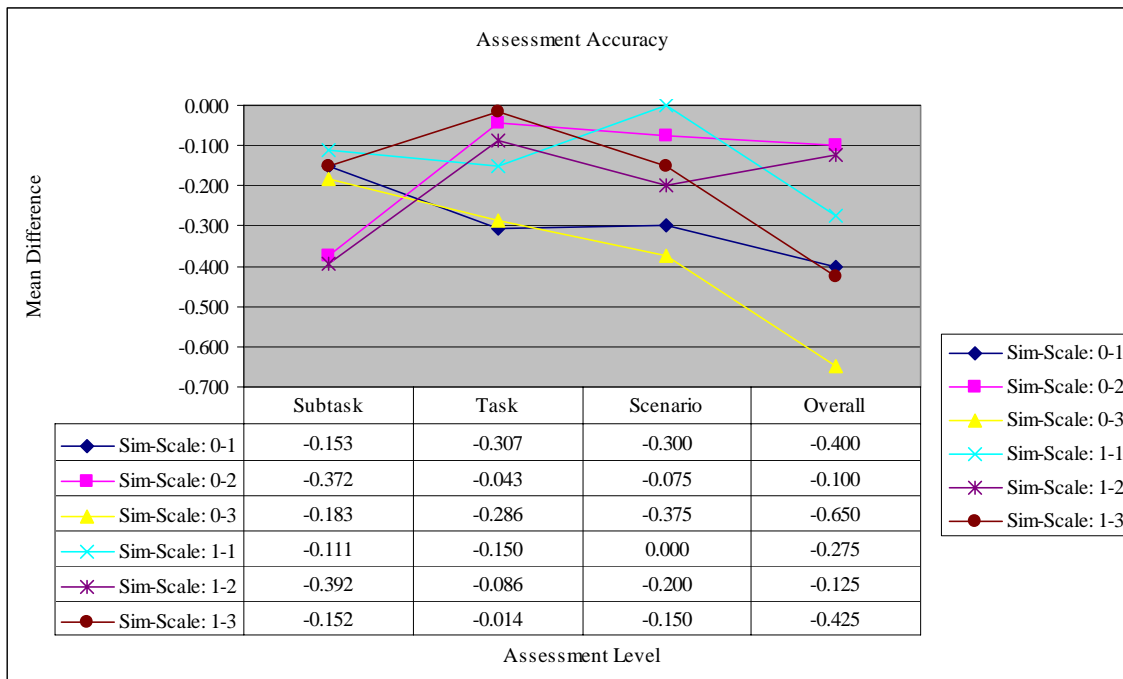


Figure N.5. Assessment Accuracy: Level

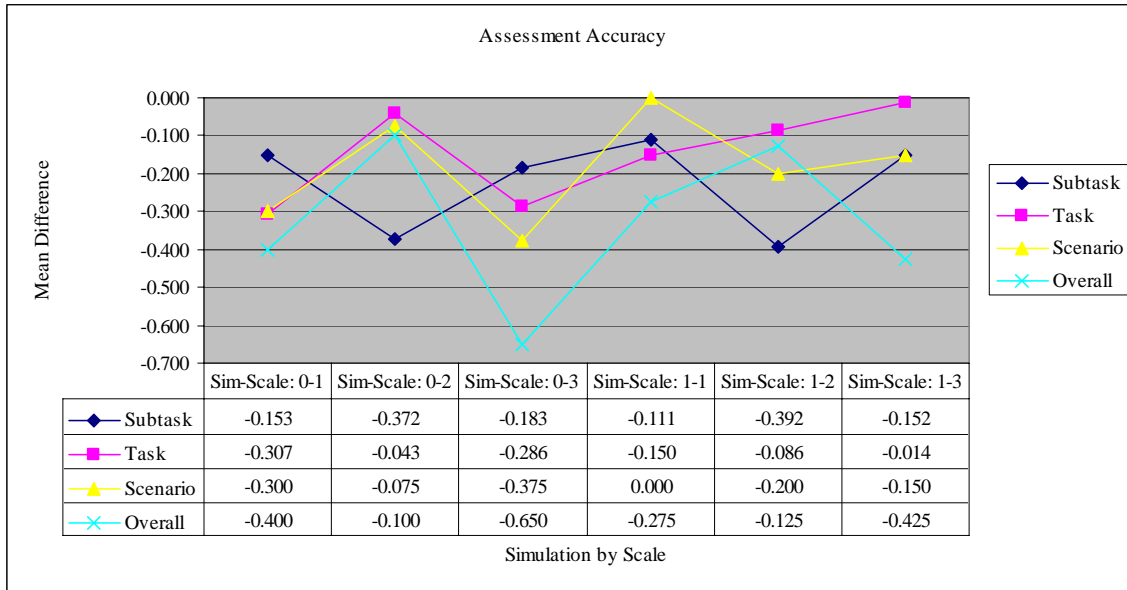


Figure N.6. Assessment Accuracy: Simulation by Scale

4. ACCURACY IMPACT

Table N.11. Ordinal Logistical Fit for Normalized Accuracy Impact Scores

Level	Prob>ChiSq			
	Whole Model Test	Effect Likelihood Ratio Test		
		Scale	Simulation Belief	Scale cross Simulation Belief
Subtask	0.0001	0.0000	0.0006	0.0101
Task	0.0001	0.0000	0.0024	0.0029
Scenario	0.0001	0.0000	0.0629	0.0381
Overall	0.0001	0.0000	0.3074	0.1216

5. EFFECT OF BIAS REMOVAL

Table N.12. Normalized, Mean Overall Assessment Scores - Minus Performance Bias

ID		Number of SMEs	Mean (Normalized 0-1 Responses)			
Simulation Belief	Scale		Overall 1	Overall 2	Overall 3	Overall 4
0	1	372 (1-2)/36 (3-4)	<i>0.583</i>	<i>0.598</i>	<i>0.540</i>	<i>0.552</i>
0	2	24	0.958	0.958	0.979	0.979
0	3	17	<i>0.565</i>	<i>0.576</i>	<i>0.482</i>	<i>0.471</i>
1	1	38	0.669	0.699	<i>0.594</i>	<i>0.624</i>
1	2	22	0.932	0.932	0.886	0.909
1	3	21	0.705	0.724	<i>0.648</i>	<i>0.686</i>
All Beliefs and Scales		159 (1-2)/158 (3-4)	0.723	0.737	0.676	0.693

Table N.13. Normalized, Mean Overall Assessment Scores - Minus Anchoring Bias

ID		Number of SMEs	Mean (Normalized 0-1 Responses)			
Simulation Belief	Scale		Overall 1	Overall 2	Overall 3	Overall 4
0	1	30	<i>0.586</i>	<i>0.576</i>	<i>0.533</i>	<i>0.538</i>
0	2	24	0.958	0.958	0.979	0.979
0	3	19	<i>0.484</i>	<i>0.505</i>	<i>0.495</i>	<i>0.484</i>
1	1	33 (1-2)/32 (3-4)	0.671	0.684	<i>0.607</i>	<i>0.643</i>
1	2	18	0.889	0.889	0.833	0.861
1	3	19	<i>0.653</i>	0.674	<i>0.589</i>	<i>0.632</i>
All Beliefs and Scales		143 (1-2)/142 (3-4)	0.701	0.708	<i>0.666</i>	0.682

Table N.14. Normalized, Mean Overall Assessment Scores - Minus Contrast Bias

ID		Number of SMEs	Mean (Normalized 0-1 Responses)			
Simulation Belief	Scale		Overall 1	Overall 2	Overall 3	Overall 4
0	1	36 (1-2)/35 (3-4)	<i>0.579</i>	<i>0.607</i>	<i>0.547</i>	<i>0.559</i>
0	2	24	0.958	0.958	0.979	0.979
0	3	24	<i>0.483</i>	<i>0.500</i>	<i>0.442</i>	<i>0.433</i>
1	1	39 (1-2)/38 (3-4)	0.681	0.707	<i>0.605</i>	<i>0.635</i>
1	2	23	0.891	0.891	0.848	0.870
1	3	22	0.627	0.682	0.609	0.645
All Beliefs and Scales		168 (1-2)/166 (3-4)	0.692	0.714	<i>0.657</i>	0.674

Table N.15. Normalized, Mean Overall Assessment Scores - Minus Confirmation Bias

ID		Number of SMEs	Mean (Normalized 0-1 Responses)			
Simulation Belief	Scale		Overall 1	Overall 2	Overall 3	Overall 4
0	1	22 (1-2)/21 (3-4)	0.630	0.623	1.208	1.227
0	2	21	1.000	1.000	1.000	1.000
0	3	14	0.457	0.457	0.386	0.386
1	1	22	0.747	0.740	0.649	0.662
1	2	21	0.905	0.929	0.857	0.881
1	3	18	0.600	0.644	0.622	0.622
All Beliefs and Scales		118 (1-2)/117 (3-4)	0.741	0.750	0.704	0.714

Table N.16. Normalized, Mean Accuracy Impact Score Difference – Results Without Bias Minus All Results

	Change in Means Accuracy Impact Scores						
	<i>Mean (total)</i>	<i>Sim-Scale: 0-1</i>	<i>Sim-Scale: 0-2</i>	<i>Sim-Scale: 0-3</i>	<i>Sim-Scale: 1-1</i>	<i>Sim-Scale: 1-2</i>	<i>Sim-Scale: 1-3</i>
Subtask	-0.0333	0.0330	-0.0092	-0.0304	0.0095	-0.0154	-0.0248
Task	0.0503	0.0604	0.0004	0.0580	0.0281	0.0522	-0.0800
Scenario	0.0939	0.1315	0.0072	0.0437	0.0644	0.0404	0.0267
Overall	0.1839	0.1022	0.0833	0.1369	0.2875	0.1030	-0.0400
Total	-0.0243	0.0365	-0.0073	-0.0216	0.0157	-0.0090	-0.0265

Table N.17. Percentage Change in Normalized, Mean Accuracy Impact Score Difference – Results Without Bias Minus All Results

	% Change in Means Accuracy Impact Scores						
	<i>Mean (total)</i>	<i>Sim-Scale: 0-1</i>	<i>Sim-Scale: 0-2</i>	<i>Sim-Scale: 0-3</i>	<i>Sim-Scale: 1-1</i>	<i>Sim-Scale: 1-2</i>	<i>Sim-Scale: 1-3</i>
Subtask	-14.79%	22.29%	-2.41%	-11.97%	8.67%	-3.91%	-12.57%
Task	28.49%	21.89%	0.87%	17.95%	15.91%	57.82%	-87.50%
Scenario	45.02%	53.07%	13.09%	9.24%	76.66%	26.67%	9.09%
Overall	49.41%	22.92%	100.00%	19.33%	82.14%	75.56%	-8.70%
Total	-10.81%	22.92%	-2.05%	-8.02%	13.61%	-2.40%	-13.41%

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX O. ILLUSTRATION OF FACE VALIDATION SHORTCOMINGS

To illustrate a major shortcoming in the current process of cognitive-model validation—namely, the face validation of overt behaviors without considering the cognitive process behind them—this section explores a tactical movement-and-navigation scenario involving a company of dismounted infantry moving north through a mountain range in Korea. The objective is to seize an airstrip nestled in the hills. An enemy infantry platoon is known to be entrenched in a battle position (BP) in the hills directly north of the airstrip (BP_01). Based on information from friendly higher headquarters, two enemy observation posts (OPs) are templated in the hills to the southwest (Figure O.1). These OPs overlook the high-speed avenues of approach to the west and south of the airstrip. These positions are most likely manned by two or three soldiers with radios, small arms, binoculars, and night-vision devices. According to the battalion intelligence officer (S2), the enemy has only had time to emplace protective obstacles around its defensive positions.

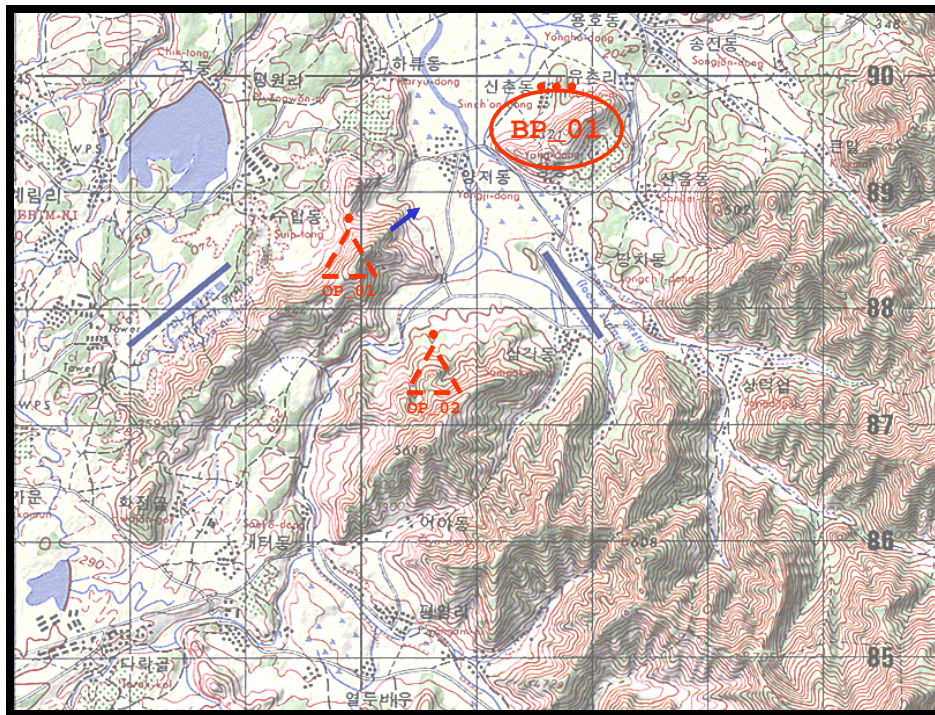


Figure O.1. Templated Enemy Situation

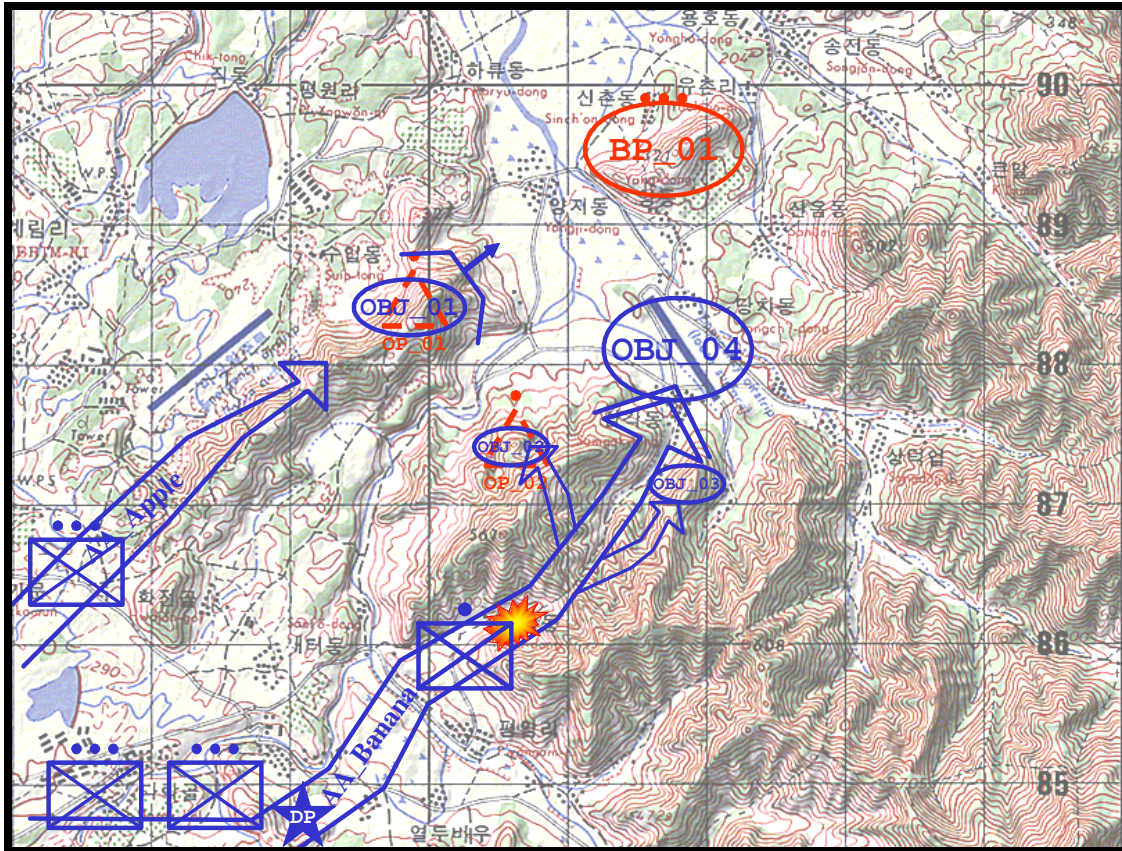


Figure O.4. Enemy Action

Most cognitive models are robust enough to handle such a situation by providing a course of action. The commander could continue forward with his forces and take up a defensive position once he his main body comes under effective fire. Continuing with the original plan is a solution routinely found in rule-based models such as Soar or COGNET. Since exhaustive examination of all possible situations is impractical in the non-linear modern-day battlefield, it is unlikely that all possible situations could be foreseen and codified in a rule set. As a result, more elegant solutions may not be presented as options in a simulation.

Is continuing with the original plan a valid solution? Depending on how a SME believes the commander comes to his conclusion, it may certainly be seen as valid. If the commander knows the enemy is waiting in the western pass (Alternative 3, Figure O.5), he may discount this option. If he believes the terrain to the east (Alternative 2, Figure O.5) is too rugged to allow him to reach the main objective by 0500, he may reject that

option. If he feels he must move forward to retrieve the men of his initial scouting force, he may see the original route as his best option. But if all we know is that he has continued to move his forces along AA_Banana, it is impossible to determine whether the decision was based on a valid thought process.

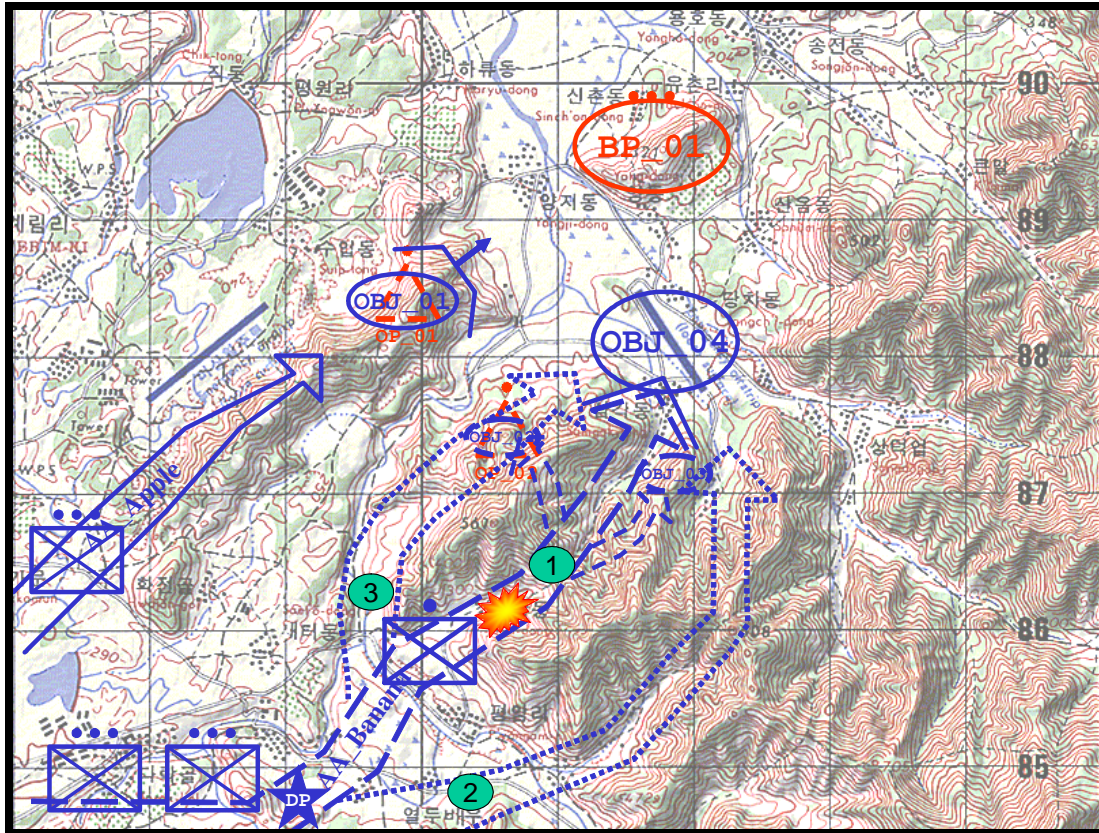


Figure O.5. Friendly Forces Potential Decisions

Pressing along the original route in the face of environmental or tactical alternations, solely because that was the original or scripted plan, is unrealistic and invalid. It is often said that an operations order never survives first contact. Most operations orders are used as a foundation from which to make adjustments. A good order provides leaders with a clear indication of the commander's desired end status and, at the same time, gives ground commanders the flexibility to execute the mission as they judge necessary. A leader who follows orders to the letter without accounting for the exigencies of the battlefield may lose both battle and unit.

SMEs, when faced with limited information, may give a cognitive model the benefit of the doubt and declare the behaviors of entities on the synthetic battlefield valid; but they may view the same actions as invalid when evaluating actual forces on the ground, because they may more closely consider the information available and the leader's decision.

A commander choosing any of the three routes in Figure O.5 might be making a valid decision. He might be making an invalid decision. The verdict depends on the facts at hand, the commander's intent, and his reason for choosing the route. He may choose a viable solution (leading to overt action) but his reasoning may be inappropriate or invalid. Determinations regarding his cognitive processes cannot be made by observing overt behavior alone.

APPENDIX P. KEY PLAYERS IN VERIFICATION, VALIDATION AND ACCREDITATION

Table P.1 outlines the roles of key players in the DoD modeling and simulation VV&A process. This table is excerpted from DMSO's VV&A Recommended Practices Guide Reference Document, "Key Concepts of VV&A".

Table P.1. Typical Roles and Responsibilities Associated with Modeling and Simulation Verification, Validation and Accreditation From [DEPA 01e]

Role Activity	User	M&S PM	Developer	V&V Agent	Accreditation Agent	SME
Define Requirements	Lead	Monitor	Assist	Review	Review	Assist
	Approve					
Define Measures	Lead	Monitor	Assist	Assist	Assist	Assist
	Approve					
Define Acceptability Criteria	Assist	Monitor	Assist	Assist	Lead	Assist
	Approve					
Plan M&S Development or Modification*	Assist	Lead	Assist	Assist		
		Approve				
Develop V&V Plans	Review	Assist	Review	Lead	Assist	
		Approve				
Develop Accreditation Plan	Review	Assist		Assist	Lead	
	Approve					
Verify Requirements	Lead-alt	Monitor	Assist	Lead	Assist	Assist
	Approve					
Develop Conceptual Model**	Assist	Monitor	Lead			Assist
	Approve					
Validate Conceptual Model	Assist	Monitor	Assist	Lead		Assist
	Approve					
Develop Design***		Monitor	Perform			
		Approve				
Verify Design	Approve	Monitor	Assist	Lead		Assist
Implement Design		Monitor	Perform			
		Approve				
V&V Data	Approve	Monitor	Assist	Lead		Perform
Verify Implementation	Approve	Monitor	Assist	Lead		Assist

	Role Activity	User	M&S PM	Developer	V&V Agent	Accreditation Agent	SME
	Test Implementation	Approve	Monitor	Lead	Assist		Assist
	Validate Results	Assist	Monitor	Assist	Lead		Assist
		Approve					
	Prepare V&V Report				Perform		
	Configure for Use	Assist	Lead	Assist			
			Approve				
	Gather Additional Accreditation Info	Monitor	Assist		Assist	Lead	Assist
	Conduct Accreditation Assessment	Monitor				Perform	Assist
	Prepare Accreditation Assessment Report					Perform	
	Determine Accreditation	Perform					
	Prepare Accreditation Report					Perform	
	Lead Leads the task. Normally involves active participation from others						
	Perform Actually does the task. Normally involves little active participation from others						
	Assist Actively participates in task (e.g., conducting tests, providing information)						
	Review Participation normally limited to reviewing results of task and providing recommendations						
	Monitor Oversees task to ensure it is done appropriately but does not normally participate						
	Approve Determines when an activity is satisfactorily completed and another can begin. Determines what activity should be pursued next (e.g., whether to continue on to the next scheduled activity or to return to a previous activity).						
*This activity refers to planning and scheduling of any M&S development, modification, or preparation							
**This activity refers to development of new as well as modification of existing conceptual models							
***This activity refers to development of new M&S designs as well as modification of existing M&S designs							

APPENDIX Q. VALIDATION PLAN

Appendix Q is the validation plan used for the initial research study. It is arranged based on the format provided by NAVMSO and therefore, does not match the format of the remainder of this document.

Validation Plan

Human Behavior Representation Validation Plan

Version 1.3

March 15, 2004

1. BACKGROUND

Research is currently underway at the Naval Postgraduate School to study the effects of bias on the assessment of model performance. The purpose of the research project is to determine if research personnel can identify and mitigate subject matter expert (SME) bias in the process of face validation in order to enhance face validation results. Thus, the focus of this plan is to provide basic information for the face validation of a human behavior representation (HBR) model. The validation agent used the Verification and Validation (V&V) Plan distributed by the Navy Modeling & Simulation Management Office in the Verification, Validation, and Accreditation (VV&A) Documentation Tool to format the plan.

1.1. M&S Description

The model used for this research is an agent-based model known as Map Awareness Non-uniform Automata (MANA). In the context of this research, research personnel use MANA to provide visual display of human behaviors for individual dismounted soldiers. The following description of MANA is drawn directly from the *MANA, Map Aware Non-uniform Automata, Version 3.0, Users Manual (Draft)*.

The Defence Technology Agency of New Zealand developed MANA to conduct research into the implications of chaos and complexity theory for combat and other military operational modeling. MANA is in the general class of models known as Agent-Based Models (ABM) and developed based on the earlier model Irreducible Semi-Autonomous Adaptive Combat (ISAAC), and its follow-on, Enhanced ISAAC Neural Simulation Toolkit (EINStein) created by the Center for Naval Analyses.

As with many ABMs, MANA contains entities controlled by decision-making algorithms. Specifically, MANA contains entities representing military units that make their own decisions, as compared to rule-based models many of which have behaviors that are explicitly determined in advance by the programmers or model developers. MANA model uses a “memory map” to provide entities with goals, which guide them about the battlefield.

MANA is further classified as a Complex Adaptive Systems (CAS). There are many aspects of MANA, which allows its classification as a CAS. Some of these general characteristics are:

- The model has the ability to exhibit “global” behavior, which materialize based on local interactions.
- The model uses the process of feedback to update agents on changes to the environment.
- Similar to a Neural Network, one cannot analyze the model by decomposing it into simple independent parts.
- Similar to human behavior, agents interact with each other in a non-linear manner, and “adapt” to their local environment.

MANA has the ability to incorporate several characteristics, which ISAAC did not have when MANA was initially developed. These include:

- Group memory of enemy contacts provides agents with situational awareness. MANA uses two mechanisms to provide situational awareness, “squad map” and “inorganic map”. The “squad map” maintains squad contacts. The “inorganic map” stores contacts based on communications from other units.
- Communications exists between units in order to pass contact information. This information can be made imperfect based on the loss of communications based on imperfect communications influenced by unit activities and environmental conditions.
- Terrain Maps are integrated which contain features such as roads which agents can follow to increase speed and undergrowth that agents can use for concealment.
- The use of waypoints for general routes may enhance movement. The waypoints provide intermediate goals to facilitate coordination of units and achievement of an ultimate goal.
- Agent personalities can be event-driven. Events (e.g., making enemy contact being shot at, engaging others, reaching a waypoint, etc.) can activate a special personality trait, which last for a certain time or until modified by another event. Personality changes can be set for individuals or an entire unit.

MANA divides its parameters into four categories: personality weightings, move constraints, basic capabilities, and movement characteristics. Personality weightings, determine an automaton’s propensity to move towards friendly or enemy units, towards its waypoint, towards easy terrain, and towards a final goal point. Next, move constraints act as conditional modifiers. An example of a modifier is the “Combat” parameter, which determines the minimum local numerical advantage a group of agents needs before the unit approaches the enemy. The third set of parameters describes the basic capabilities of the agent with respect to its use of weapons, its use of sensors, its movement speed, and

its tendencies for interaction with other agents. The final set of parameters provides options on the movement characteristics of the agents, including the effects of terrain on agent speed, the degree of random agent movement, and if agents attempt to avoid obstacles [GALL 2003].

1.2. M&S Program Objectives

The objective of this research is to identify possible SME bias in the face validation of an HBR in an ABM and identify possible ways to mitigate the bias, if possible. Face validation is the use of SMEs who observe the results of model behaviors in the context of prescribed scenarios and make a decision whether the behaviors meet a user's¹⁰⁴ needs for realism. This is normally a qualitative determination [VERI 2001]. For this plan, the user and the validation agent are the same person.

The SMEs assess the performance of the model in an urban environment to determine if the human behaviors reasonably replicate individuals and squads performing the prescribed tasks. Squads are groups of nine to eleven individuals working towards a common goal.

1.3. Development Reference Materials

The following reference material supports this V&V plan.

ARTEP 7-8-MTP: Mission Training Plan for the Infantry Rifle Platoon and Squad. (Mission Training Plan)(2001). Washington, DC: Headquarters, Department of the Army.

Balci, O. (1998). Verification, Validation, and Testing. In J. Banks (Ed.), *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice* (pp. 335-393). New York: John Wiley & Sons: Co-published by Engineering & Management Press.

Carley, K. M. (1996). *Validating Computational Models* (No. United States Navy Grant No. N00014-93-1-0793 (UConn FRS 521676)): Office of Naval Research (ONR).

Department of Defense (DoD) 5000.59-P: Modeling and Simulation (M&S) Master Plan. (1995). Alexandria, VA: Under Secretary of Defense for Acquisition and Technology, Department of Defense.

Department of Defense Directive (DoDD) 5000.59: Department of Defense Modeling and Simulation (M&S) Management. (1994). Alexandria, VA: Department of Defense.

¹⁰⁴ A user is any individual or organization who will be utilizing the model for a specific research, study, or training purpose.

- Department of the Army Pamphlet: DA Pam 5-11: Verification, Validation, and Accreditation of Army Models and Simulations.* (1999). Washington, DC: Headquarters, Department of the Army.
- Department of the Army Regulation: Army Regulation (AR) 5-11: Management of Army Models and Simulations.* (1997). Washington, DC: Headquarters, Department of the Army.
- FM 3-06.11: Combined Arms Operations in Urban Terrain.* (2002). Washington, DC: Headquarters, Department of the Army.
- Harmon, S.Y. (ed.). (16 December 1998). "Fidelity ISG Glossary," Ver 3.0. Simulation Interoperability Standards Organization (SISO), Fidelity Implementation Study Group (ISG), [WWW Document]. http://www.sisostds.org/doclib/doclib.cfm?SISO_RID_1000789 (viewed 02 July 2002).
- Galligan, D. P., Anderson, M. A., & Lauren, M. K. (2003). *MANA, Map Aware Non-uniform Automata, Version 3.0, Users Manual (Draft)*. Unpublished manuscript.
- Goerger, S. R. (2003, 20 - 24 July). *Validating Human Behavioral Models for Combat Simulations Using Techniques for the Evaluation of Human Performance*. Paper presented at the 2003 Summer Computer Simulation Conference, Montreal, Quebec, Canada.
- Gonzalez, A. J., & Murillo, M. (1999, 14-19 March). *Validation of Human Behavior Models*. Paper presented at the 1999 Spring Simulation Interoperability Workshop (SIW), Orlando, FL.
- Harmon, S. Y., & Metz, M. L. (2001). *Characterizing SME Referents: An Example of Objective Validation* (Power Point Presentation): ZETETIX.
- Harmon, S. Y., Hoffman, C. W. D., Gonzalez, A. J., Knauf, R., & Barr, V. B. (2002, 22-24 October). *Validation of Human Behavior Representations*. Paper presented at the Foundations '02, a Workshop on Model and Simulation Verification and Validation for the 21st Century, Kossiakoff Conference & Education Center, Johns Hopkins University Applied Physics Laboratory, Laurel, MD.
- Office of the Director of Defense Research and Engineering (DDR&E), & Defense Modeling and Simulation Office (DMSO). (2001). *Department of Defense Instruction (DoDI) 5000.61 (Draft): Department of Defense Modeling and Simulation (M&S) Verification, Validation, and Accreditation (VV&A)*. Washington, DC: Under Secretary of Defense for Acquisition and Technology, Department of Defense.
- Sargent, R. G. (1979). *Verifying and Validating Simulation Models*. Paper presented at the 11th Conference on Winter Simulation, San Diego, CA.
- Sargent, R. G. (1999). *Verifying and Validating Simulation Models*. Paper presented at the 31st Conference on Winter Simulation: Simulation---A Bridge to the Future, Phoenix, AZ.

Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide (RPG): Reference Document - A Practitioner's Perspective on Simulation Validation and Conceptual Model Development and Validation. (2001, 15 August). Retrieved 24 January 2003, from http://www.msiac.dmsomil/vva/ref_docs/val_lawref/default.htm#toc1

Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide (RPG): Special Topic - Validation. (2000, 15 August). Retrieved 24 January 2003, from http://www.msiac.dmsomil/vva/Special_topics/Validation/Validation.htm

Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide (RPG): Special Topic - Validation of Human Behavior Representations. (2001, 25 September). Retrieved 24 January 2003, from http://www.msiac.dmsomil/vva/Special_topics/hbr-Validation/default.htm

VV&A Documentation Tool; Automating the VV&A Documentation. (2003). Navy Modeling and Simulation Office (NAVMSMO).

1.3.1. Conceptual Model

The conceptual model is not within the scope of this research.

1.3.2. Configuration Management

The configuration management is not within the scope of this research.

1.3.3. Data

The validation agent derives referent¹⁰⁵ material from *ARTEP 7-8-MTP: Mission Training Plan for the Infantry Rifle Platoon and Squad* and from SMEs. The data is limited to the subtasks in support of three distinct tasks the SMEs will assess:

- Conduct Tactical Movement in a Built-up Area (Infantry Squad) (Task 07-3-1279); 21 subtasks
- React to Snipers (Infantry Squad) (Task 07-3-1406); 23 sub tasks
- Conduct a Strongpoint Defense of a Building (Infantry Squad) (Task 07-3-1162); 15 subtasks

FM 3-06.11: Combined Arms Operations in Urban Terrain and course instruction at the Infantry Captains Career Course proved the specific explanation, examples of proper performance for the tasks and subtasks.

1.4. V&V History

¹⁰⁵ Referent is the “codified body of knowledge about a thing being simulated.” [HARM 98] In the case of HBR and this research, this would consist of at least one of the six levels of correspondence. Referent is the best information we have about the simulated objects functionality and performance. The referent provides the standards against which to compare the results of models and simulations to assess the level of fidelity they are able to replicate. [VERI 00]

<i>Status</i>	Human Behavior Representation Validation Plan
<i>Reason for Change</i>	Review by outside validation agents and policy developers
<i>Description</i>	The Human Behavior Representation Validation Plan is currently under review to ensure completeness and applicability.

2. OBJECTIVES

2.1 Intended Use

The intended use of the model is to provide analytical insight into the non-linear nature of small unit interactions in an urban environment. This includes entity level interactions at individual soldier to squad level (nine to ten personnel per side). The model provides an environment to view and assess the performance of dismounted soldiers, teams, and leaders performing the tasks listed in Section 2.3 Data Requirements.

2.2. M&S Requirements

The primary task of this research is to review face validation procedures and provide insight into the issues regarding the use of SMEs. SMEs view the human behavior representation through an agent-base model. The SMEs view three separate scenarios and assess the performance of the entities in the simulation based on three tasks with their associated subtasks as described in *ARTEP 7-8-MTP: Mission Training Plan for the Infantry Rifle Platoon and Squad*. SMEs assess tasks and subtasks in accordance with the standards set forth in *FM 3-06.11: Combined Arms Operations in Urban Terrain* and instruction at the Infantry Captains Carrier Course.

2.3. Data Requirements

The data required for this validation effort is limited to the subtasks in support of three distinct tasks outlined in Section 1.3.3 Data:

- Conduct Tactical Movement in a Built-up Area
- React to Snipers
- Conduct a Strongpoint Defense of a Building

The validation agent and SMEs assess the data within the context of urban operations for dismounted soldiers at the individual and squad level.

2.4. M&S Assumptions and Limitations

Due to their limited exposure time to the model, SMEs are provided a narrow scope of focus, three dismounted soldier tasks in an urban environment. The terrain box

is limited to 400 meters by 400 meters. The small size of the play box for this research prevents using indirect fire assets, mortars, artillery, etc., as well as smoke. Although the model can handle more entities, to maintain SME focus, the combatant forces in view are two squad sized elements (one offensive and one defensive). The validating agent prepares scenarios to allow for the preparation of the battlefield by supporting forces. Civilians are present in the environment to provide additional elements for the forces to identify and track. The simulation is two-dimensional and thus MANA does not play multiple story buildings and sub terrain features.

As a distillation model, MANA is designed to create a bottom-up abstraction of a scenario, which compares the essence of a situation while avoiding unessential detail that could potentially limit the amount and type of information provided to the SME. Limiting information availability could inhibit SME insight into the thought process the agent used to choose or not choose a specific action.

The model plays multiple forces grouped into three general categories: friend, foe, or neutral.

3. VALIDATION MANAGEMENT STRATEGY

3.1. Validation Approach

The validation approach for this research consists of the face validation of three tasks as outlined in Section 1.3.3 Data. This approach requires the validation of the referent material and the behaviors of the model.

3.1.1. Validate Data

The validating agent identified and collected referent based on the limited focus of the research to three tasks: Conduct Tactical Movement in a Built-up Area, React to Snipers, and Conduct a Strongpoint Defense of a Building. Research personnel use *ARTEP 7-8-MTP: Mission Training Plan for the Infantry Rifle Platoon and Squad* to define the tasks and subtasks. *FM 3-06.11: Combined Arms Operations in Urban Terrain* is used to identify the standards for the proper performance of these tasks and subtasks. Subject matter experts assess the performance of the model based on its ability to maintain general doctrinal correctness.

Research personnel scrub the subtasks for viability and applicability based on the responses of five pilot study personnel. Although some subtasks are not applicable

based on the scenario, they remain on the assessment sheet to maintain uniformity of the forms across scenarios.

3.1.2. Validate Results

Validation results are based on SME face value assessments of the models behaviors. The overall assessment is based on the execution of three scenarios, which encompass three distinct tasks performed seven times and 146 subtasks. Elements assessed as valid by two thirds or more of the participants, are said to be valid. The validation agent deems the model valid for the needs of the user, with respect to the three assessed tasks, if two thirds or more of the SMEs' overall assessments classify the model as valid. The overall assessment is based on the SMEs' assessments of the model's execution of three scenarios, which encompass three distinct tasks performed seven times. There are 146 subtasks, which support the seven tasks.

The validating agent categorizes SME observations into discrete and qualitative categories. SMEs provide categorical qualitative responses for each subtask, task, scenario, and for the overall assessment of the HBR performance utilizing a 7-Point Likert Scale. SMEs use a 7-Point Likert Scale to provide room for more flexible responses. The scale values are as follows:

- 0 – Not applicable (NA)** or no means of determining
- 1 – Strongly agree** the task, step, or performance measure was *improperly* performed
- 2 – Agree** the task, step, or performance measure was *improperly* performed
- 3 – Not sure** but tend to agree the task, step, or performance measure was *improperly* performed
- 4 – Undecided**
- 5 – Not sure** but tend to agree the task, step, or performance measure was *properly* performed
- 6 – Agree** the task, step, or performance measure was *properly* performed
- 7 – Strongly agree** the task, step, or performance measure was *properly* performed

Assessment scores of one, two, and three are “No-Go” responses while assessment scores of five, six, and seven are “Go” responses.

The validation agent compares SME results for a given subtask, task, scenario, and overall assessment for inter SME consistency. Where variance exists, the

validation agent seeks clarification from SME comments for the given assessment question in an attempt to resolve the variance.

For consistent results, the validation agent calculates the mean score for each level of assessment in order to determine the validity of each element and for the overall validity score of the tasks performed by the model. The validation agent bases results on the overall validity score of the assessed elements. Those items assessed as “Go” (assessment scores of five, six, or seven) by two thirds or more of the participants, are said to be valid. The validation agent deems the model valid for the needs of the user, with respect to the three assessed tasks, if two thirds or more of the SMEs overall assessments classify the model as valid.

The validation agent surveys those elements identified as not performed to standard to gather specific input for possible model enhancements.

3.2. Points of Contact

Points of contact (POCs) are the primary representatives for each aspect of the validation process, which are within the scope of the research.

3.2.1. V&V Agent

MAJ Simon Goerger
MOVES, Naval Postgraduate School
700 Dyer Road, Room 265, Naval Postgraduate School
Monterey, California 93943-5001
Phone: (831) 656-3733
DSN: 756-3733
Fax: (831) 656-7590
E-Mail: srgoerge@nps.navy.mil

3.2.2. M&S Developer

Defence Technology Agency of New Zealand

3.2.3. Subject Matter Expert(s)

ICCC Students
Infantry Captains Career Course
6751 Constitution Loop (Bldg 4)
Fort Benning, Georgia 31905

3.2.4. M&S User(s)

MAJ Simon Goerger
MOVES, Naval Postgraduate School

700 Dyer Road, Room 265, Naval Postgraduate School
Monterey, California 93943-5001
Phone: (831) 656-3733
DSN: 756-3733
Fax: (831) 656-7590
E-Mail: srgoerge@nps.navy.mil

3.2.5. Data POC

MAJ Simon Goerger
MOVES, Naval Postgraduate School
700 Dyer Road, Room 265, Naval Postgraduate School
Monterey, California 93943-5001
Phone: (831) 656-3733
DSN: 756-3733
Fax: (831) 656-7590
E-Mail: srgoerge@nps.navy.mil

3.3. Validation Program Control

The scope of this plan is the face validation of the agent-based model's ability to perform the three prescribed tasks. The following describes the procedures for the execution of the face validation of the tasks described in Section 1.3.3 Data.

The face validation of the model is broken down into two segments: data collection and assessment of results. The data collection segment, which includes the use of SMEs, is broken into three phases. The validating agent performs an assessment of results at the conclusion of the data collection segment (see Section 3.1.2 Validate Results).

The validation agent selects SMEs based on their training in the area of urban operations, their experience with training squads and assessing individual soldiers and teams in the execution of infantry tactics, and have at least seven years experience with military doctrine and operations, and three years operational experience. All SMEs are from a pool of officers in the Infantry Officers Career Course. They range in age from 25 to 41 and all SMEs possess at least a bachelor's degree.

Once selected to participate in the face validation process, SMEs undergo three distinct phases to prepare them for the data collection portion of the face validation process: familiarization, training, and data collection.

- *Familiarization*: Each SME undergoes a familiarization phase covering a series of briefings designed to a) explain the objectives of the experiment, b) review appropriate doctrine and tactics, techniques, & procedures (TTPs) for the task(s) and subtasks to be assessed, c) acquaint participants with validation techniques, and d) review response forms.
- *Training*: SMEs undergo a training scenario to expose them to the material they are introduced to in the familiarization phase and to ensure their proficiency at the assessment process. The training scenario is 90 seconds in length and displays obviously, but subjectively questionable “poor” and “good” performance of defense of a building.
- *Data Collection*: During the data collection phase, SMEs validate CGF performance for a series of tactical military tasks displayed in the MANA view frame. The scenarios are 90 seconds in length and display performance of movement and reconsolidations tasks for an urban environment. They begin by recording the method they feel would be viable for the execution of the mission. SMEs then watch a fast forward run of the scenario prior to a run of the scenario at near real time (slower) speeds. SMEs record their observations and opinions at predetermined points in the scenarios. Each participant is asked to assess behaviors in two offensive and one defensive scenario.

The two offensive scenarios consist of two tasks: Conduct Tactical Movement in a Built-up Area and React to Snipers. Each offensive scenario starts with movement, shifts to react to sniper, and concludes with movement. The defensive scenario consists of one task, Conduct a Strongpoint Defense of a Building, which SMEs assess one time. SMEs provide categorical qualitative responses for each subtask, task, scenario, and for the overall assessment of the HBR performance utilizing a 7-Point Likert Scale. The SME assigns a score of five or greater when the SME feels the element (subtask, task, scenario, or overall question) is performed to standard in accordance with the procedures outlined in *FM 3-06.11: Combined Arms Operations in Urban Terrain*. The assumption is that the doctrinal tasks are appropriate for the needs of the user.

The validation technique asks SMEs to provide comments to amplify their categorical qualitative responses when applicable. Participants receive guidance to ensure they comment when their sublevel tally results differ from their assessment of the level. For example, when the subtask scores indicate a task is performed to level five but the SME feels the task was performed at three, the SME should provide a reason for the inconsistency of subtask mean to task score.

The overall assessment is based on the execution of three scenarios, which encompass three distinct tasks performed seven times and 146 subtasks. Elements assessed as valid (assessment scores of five, six, or seven) by two thirds or more of the participants, are said to be valid. The validation agent deems the model valid for the needs of the user, with respect to the three assessed tasks, if two thirds or more of the SMEs overall assessments classify the model as valid.

4. FUNDING

The Navy Modeling & Simulation Management Office (N61M), 2000 Navy Pentagon, Washington, DC 20350-2000, provides funding for the HBR validation research.

5. VALIDATION SCHEDULE

Task	Time Line
Data Assessment by SMEs	01-05 September 2003
Pilot Study	01-05 September 2003
Model Assessment by SMEs	18-29 September 2003
Analysis of Results	29 October 2003 – 05 April 2004
Reporting of Results	05 April 2004 – 18 June 2004

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX R. MODEL TAXONOMIES

Appendix Q describes five categorizations of models in use by the Department of Defense. These categories are simulation typologies, real-time versus non real-time models, hierarchy of models and simulations, military simulations versus games, and model domains. These categories are to help the reader better understand the scope of this research.

There are three *simulation typologies*: live, virtual, and constructive.¹⁰⁶ Figure 64 is a visual depiction of the relationship between the three simulation typologies. In a live simulation, cognitive processes used by human actors determine most of their behaviors. In addition, some weapon systems may have embedded cognitive models. This research does not deal with the typology of live simulations. Virtual simulations include the additional complexity of real users interacting with the simulated equipment. The interaction of real humans and cognitive models could add a bias to the base evaluation or validation of a model as humans interact with the environment. Humans sometimes perform non-repeatable actions. Users' reaction times change as humans learn to "play" the simulation, fatigue issues, or a user enhances his proficiency with the interface.

¹⁰⁶ Live simulations consist of people interfacing with real equipment. Virtual simulations deal with people interacting with simulated equipment. Constructive simulations are simulated people operating in a simulated environment [HUGH 97].

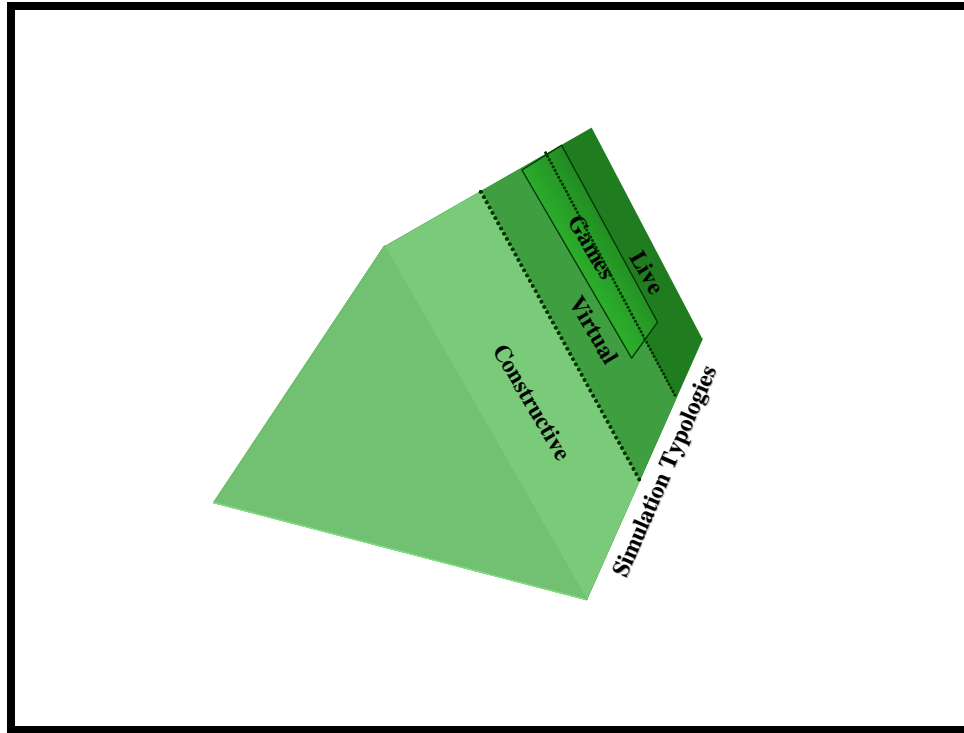


Figure R.1. Simulation Typologies From [HUGH 97] [GOER 03]

Models can be categorized based on their ability to maintain “*real-time*” performance or “*non real-time*” performance. Real-time models are usually associated with virtual simulations, as models provide human participants with timely information, reactions, and effects. Non-real time models are usually associated with constructive simulations and research personnel frequently use them for analytical studies. The use of non real-time models allows users to conduct studies using the highest fidelity algorithms available with little concern for the computational speed requirements. One method used to ensure real-time performance requires simulations to partition the limited computational power of the system to each model. This limiting of computational assets could force algorithms to provide answers that are not optimal. Another approach is to use lower fidelity, less computationally intensive, algorithms. Either method of ensuring real-time performance could add bias when validating a cognitive model.

Modifying a non real-time simulation to allow it run in real-time would be relatively trivial if the combat simulation and its associated models ran at a simulation speed faster than real-time. Model developers can synchronize the simulation clock with

real-time by adding wait events to the simulation queue when no other simulation actions are scheduled to perform. However, if the combat simulation runs slower than real-time the process is more difficult. Five possible solutions are: 1) reduce algorithm complexity; 2) reduce the complexity of the scenarios to allow the simulation to execute in real-time; 3) partition system resources to force models and algorithms to provide answers at specified times to ensure execution in real-time; 4) extend simulation capabilities to operate in a multi-threaded environment, farming processes to multiple machines; or 5) wait until hardware improvements decrease model execution times allowing them to run faster than real-time and then modify the simulation to execute at real-time.

Each of these methods for modifying simulation runtime has its drawbacks. Each method also provides an interesting area of study to identify and assess the performance of specific models and systems with limited information or resources. Though these studies may not help to validate the models, they could lead to a means of simulating and evaluating the importance of information or to the discovery of optimization techniques, which could eventually aid in the development of future models.

Department of Defense also categorizes its simulations according to the simulation's level of fidelity. Some organizations classify models into a hierarchy consisting of five levels: campaign, theater, mission, engagement, and engineering [INTR 02]. The Army Modeling and Simulation Office's publication, "Planning Guidelines for Simulation and Modeling for Acquisition Requirements and Training," describes a four level hierarchy (Figure 65) where the hierarchy combines theater and campaign into one level [AMSO 00]. Hughes refers to these four levels as campaign or theater, battle or multi-unit engagement, single engagement, and phenomenological [HUGH 97].

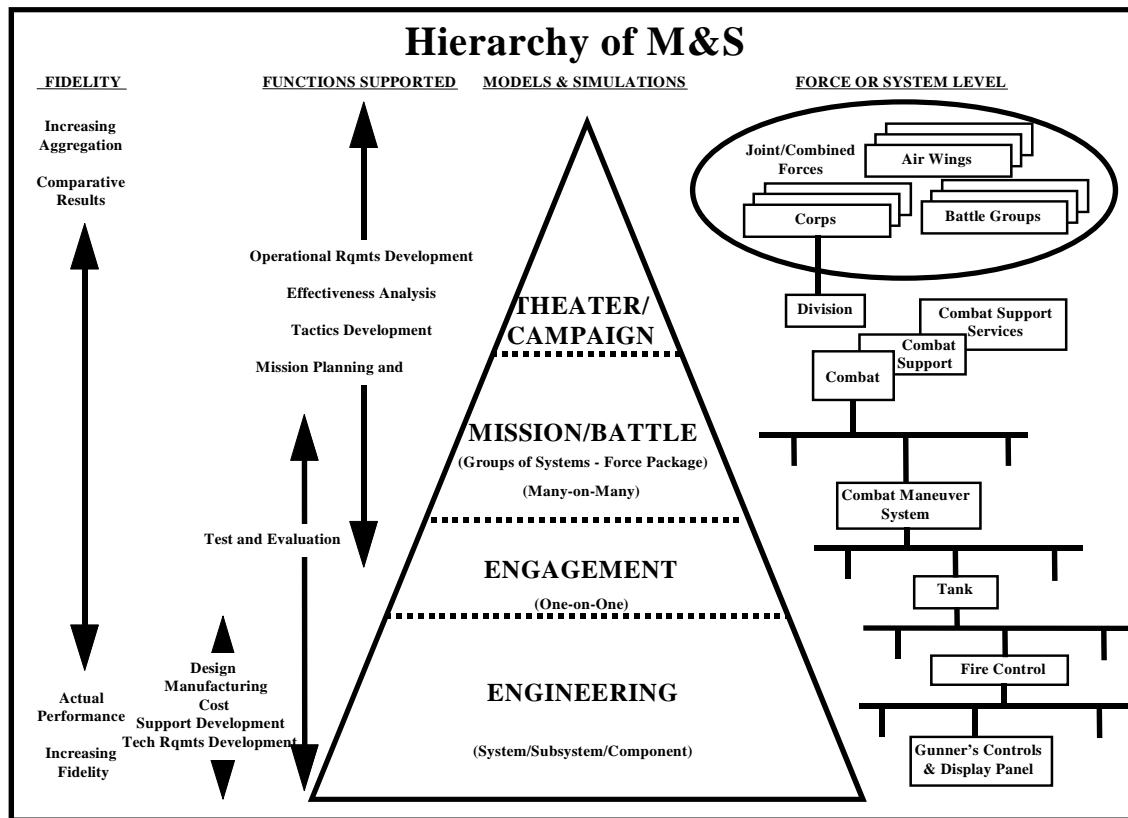


Figure R.2. Army Modeling and Simulation Office's Hierarchy of Modeling and Simulation From [AMSO 00]

This research combines the top two layers of the AMSO and Hughes' hierarchies of simulations creating a three-tier hierarchy. Figure 66 is an illustration of these three tiers: aggregate, entity, and engineering. The aggregate level consists of models that combine combat entities into groups based on functionality or association. Examples of these are a company of tanks or a brigade task force, respectively. The aggregation of forces helps reduce the computational requirements for modeling large-scale battles or scenarios. Aggregate simulations often have a cognitive component to the model, for example, the Joint Semi-Autonomous Forces simulation often aggregates forces into platoons, companies, etc., and places them under the control of a simulated "leader" who decides how the force is to be employed. The cognitive models of an aggregate model normally take into account the desires of the leadership at the level of aggregation and model the combined effects of the leaders' decisions on the subordinate elements. This

usually means less fidelity than is normally found at the entity and engineering level simulations.

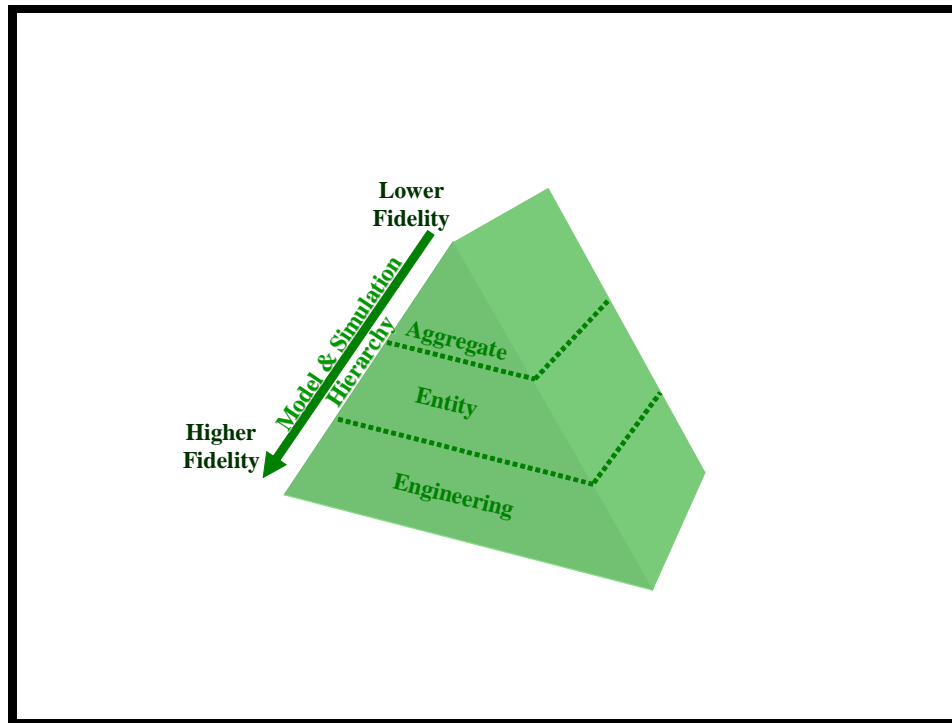


Figure R.3. Hierarchy of Modeling and Simulation From [107]

Engineering level models normally deal with one or two battlefield systems. These models may go into more detail to model the physical aspects of the system. An example of this would be modeling the hull, engine, drive train, and suspension of a tank to determine if the vehicle can operate in rough terrain. Program development offices use this form of modeling prior to building the vehicle and testing it using live simulations. Engineering level models rarely use cognitive processes, normally limiting themselves to the physics of the system.

The entity level simulations reside between aggregate and engineering level simulations. Entity level simulations represent individual platforms and the effects created by or acting on them. These models often have a behavioral component. These models focus behaviors at the individual entity level, i.e., a man, tank, etc., but can also include aggregate behaviors.

¹⁰⁷ See [HUGH 97] [AMSO 00] [GOER 03]

Numerous domains create models to provide entertainment, training, or analysis. The needs of these different communities place different constraints on the validity of the model. In contrast, for the gaming industry, the overriding concern is to manufacture a product people are interested in purchasing and playing. One can categorize games as aggregate or entity level models. Aggregate games are strategic in nature. A few examples are Axis & Allies, Risk, Fortress America, and Second Front. This research considers first shooter and role-playing games like America's Army, Tomb Raider, Jet Fighter, and Medal of Honor entity level games. No matter which category of model a game resides, game developers are not primarily concerned with realism or the accurate portrayal of human performance during the design and coding of such a model. However, although entertainment is more important to game developers than realism, the current trend in the gaming community is towards more complex and realistic human performance models.

The willingness of game developers to suspend reality in order to achieve a factor of "fun" makes the use of a gaming model for evaluation of cognitive models undesirable for this research. However, Laird and van Lent claim the gaming world is the next experimental ground for advancing research in human-level artificial intelligence [LAIR 00]. They argue that combat models are too complex and restrictive in available player roles to facilitate the study of "human-level AI". They feel the rich, interactive, real-time environment of games allows for a more robust human-computer interaction. Laird and van Lent assert that developing military simulations is expensive and time consuming, and that games can be built and modified more rapidly, facilitating the study and evaluation of human behaviors [LAIR 00] .

Laird and van Lent fail to address the issues of limited physics based modeling in games, the unrealistic capabilities games often provide users, and the possibility of users modifying their behaviors to *perform better* in simulations rather than eliciting reasonable real world human behaviors. Combat models and training simulations deal with these issues. Potentially, this makes combat models more viable than a gaming environment for studying real world human performance capabilities or behaviors.

Even if we limit our scope to DoD models, we have a large area of interest¹⁰⁸ to explore. Each service has its own set of models to cover its area of expertise. For example, the Air Force has aircraft and space simulations. The Navy has surface, subsurface, radar, sonar, weather, and aircraft models to address the needs of its missions. The Marine Corps uses a mix of surface, ground combat, and aircraft simulations. Finally, the Army deals primarily with ground combat and rotary wing aircraft simulations. Each domain has different fidelity requirements even for the same level of typology and hierarchy. An example is terrain fidelity. Fighter simulators move higher and faster requiring large areas of terrain with a relatively low density of elevation postings. Tank simulators move lower and slower requiring less area but a higher density of elevation postings. Each may be an entity level, real-time, virtual or constructive model, but each has different terrain data requirements.

Figure 67 combines these five categories into one diagram of M&S taxonomies. The taxonomy provides a means of placing into perspective the numerous areas of modeling and simulation interest and provides a pallet for describing the purpose and general capabilities of a model.

¹⁰⁸ Area of interest refers to “a geographical area from which information and intelligence are required to execute successful tactical operations and to plan for future operations.” [DEPA 01a] In this context, area of interest referees to the scenario terrain deck on which a model or simulation may operate.

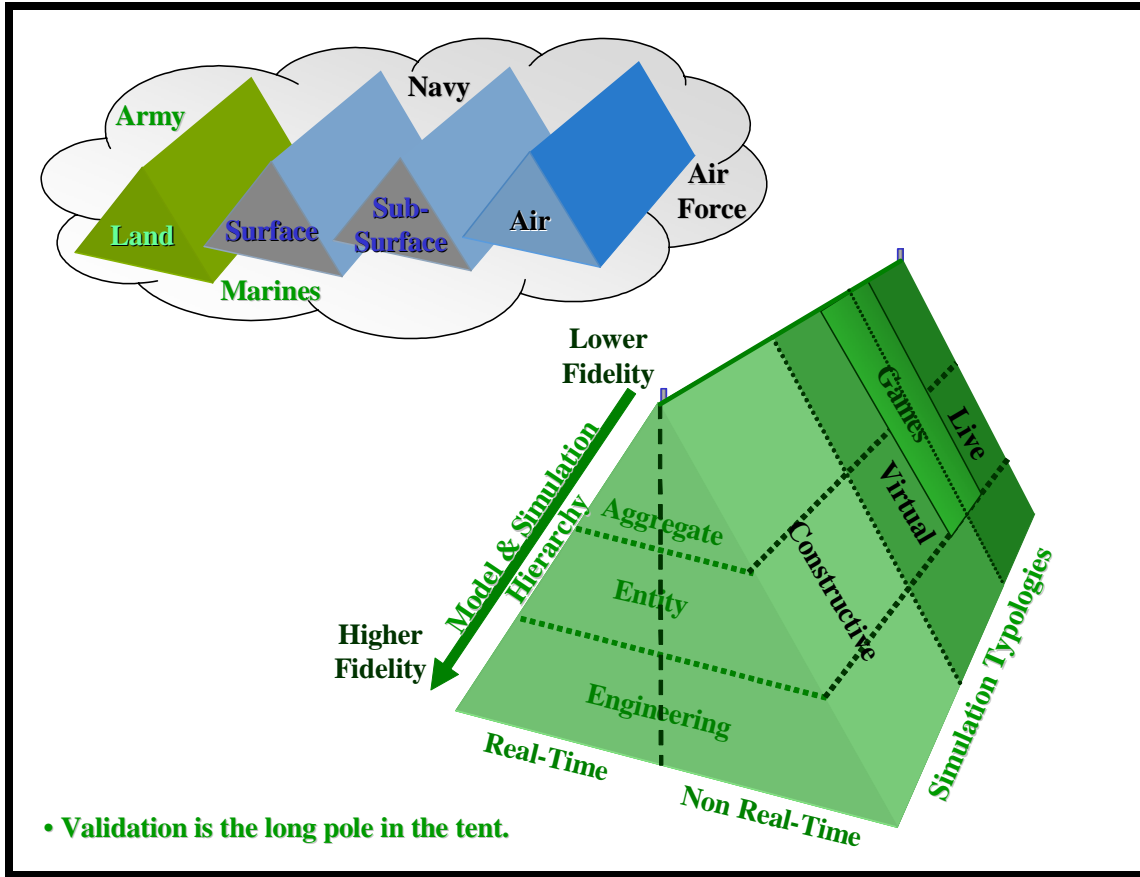


Figure R.4. Model and Simulation Taxonomies From [¹⁰⁹]

¹⁰⁹ See [HUGH '97][AMSO '00] [LAIR '00][GOER '03]

GLOSSARY

The following definitions for terms used in this dissertation are excerpted from Department of Defense Directive 5000.59, *DoD Modeling and Simulation (M&S) Management*; DMSO's *VV&A Recommended Practices Guide*, "Key Concepts;" Gary Klein's *Sources of Power; How People Make Decisions*; DMSO's *Human Behavior Representation (HBR) Literature Review*; and other DoD and professional publications.

1. Accreditation

"The official certification that a model, simulation, or federation of models and simulations and its associated data are acceptable for use for a specific purpose." [DEPA 94] This is the final stage of the verification, validation and accreditation (VV&A) process. **Accreditation** is the "official" seal of approval by the designated authority that the model is verified and valid for its intended purpose.

2. Accrediting Agent

Accrediting agents are those persons or organizations responsible for executing the verification, validation and accreditation (VV&A) process of a model, simulation, or federation and its supporting data [DEPA 01b]. The model sponsor normally designates the accrediting agent [DEPA 95].

3. Accuracy

For this research, **accuracy** is defined as the SME's average difference between the assessment key and the SME's assessment of each observation, where a difference is the assessment value from the key minus the assessment value of the SME for a given subtask, task, scenario, or overall question.

4. Accuracy Impact

For this research, **accuracy impact** is defined as the SME's average difference between the assessment key and the SME's assessment of each observation, where a difference refers to a change from Go to No-Go, Go to Unknown, No-Go to Go, No-Go to Unknown, Unknown to Go, or Unknown to No-Go.

5. Anchoring Bias

Anchoring bias emerges when an individual embraces an initial hypothesis and maintains this view regardless of incoming facts. This results in overemphasis on the hypothesis and an inappropriately minimal shift from the initial viewpoint [TVER 74] [KAHN 82] [COHE 93] [DUFF 93] [STEI 98] [PERR 03].

6. Assessment

An **assessment** or **rating** is the value (based on scale) an individual SME gives an observed model or human behavior.

7. Assessment Key

The **assessment key** is a set of subtask assessments tallied and averaged to produce tasks assessments, which when tallied and averaged produce scenario responses. The average value for the scenario responses determines the overall assessment of the behaviors. Each scale has its own assessment key and all assessment keys are consistent with each other.

8. Bias

As defined by *Webster's Dictionary*, **bias** is the “systematic error introduced into sampling or testing by selecting or encouraging one outcome or answer over others” [MERR 03].

9. Cognitive Task Analysis

A “**cognitive task analysis** is a method for capturing expertise and making it accessible for training and system design.” It results in a “... description of the expertise needed to perform complex tasks.” It consists of five steps: (1) identifying sources of expertise; (2) assaying the knowledge; (3) extracting the knowledge; (4) codifying the knowledge; and (5) applying the knowledge [KLEI 01].

10. Confirmation Bias

Confirmation bias is demonstrated when a SME overvalues select pieces of information relative to consistent evidence indicating an alternate conclusion [COHE 93] [DUFF 93] [STEI 98] [PERR 03].

11. Consistency

For this research, a SME's ability to maintain logical correspondence between the average sublevel response score and the level score is **consistency**. In other words, SMEs derive level responses logically/directly from sublevel responses.

12. Consistency Impact

For this research, the degree to which a SME's consistency/inconsistency influences the assessment of the model by changing a SME's results between sublevel and level from Go to No-Go, Go to Unknown, No-Go to Go, No-Go to Unknown, Unknown to Go, or Unknown to No-Go is **consistency impact**. In other words, does the inconsistency, when present, make a practical difference in the outcome of the assessment.

13. Contrast Bias

Contrast bias materializes when one seeks information to contradict an original hypothesis, ignoring or undervaluing evidence in support of the hypothesis [TVER 74] [KAHN 82] [PERR 03].

14. Correspondence

Correspondence is "the agreement of things with one another" [MERR 02]. In the validation domain, this term is used to describe the agreement of a model to different levels of abstraction. There are at least six levels of correspondence used in HBR validation: computational, domain, physical, physiological, psychological, and sociological [DEPA 01d].

15. Credibility

Credibility is "the relevance that the user sees in a model and the confidence that the user has that a model or simulation can serve his purpose" [DEPA 01b].

16. Decision Bias

According to Cohen, **decision bias** is "a systemic flaw in the internal relationships among a person's judgments, desires, and /or choices" [COHE 93].

17. Evaluation

Evaluation is a means of determining how well a model agrees with the portion of the real world it is simulating. It is a less stringent means of agreement than validation

and is usually based on qualitative versus quantitative data. It is used to assess the model's quality when a model is non-predictive or incapable of validation [HODG 92].

18. Fidelity

“The degree to which a model or simulation reproduces the state and behavior of a real-world object or the perception of a real-world object, feature, condition, or chosen standard in a measurable or perceivable manner; a measure of the realism of a model or simulation; faithfulness. *Fidelity* should generally be described with respect to the measures, standards, or perceptions used in assessing or stating it” [HARM 98]. The higher the model's *fidelity*, the more it corresponds to the complexities and represents the real-world element it is simulating. This term is qualitative in nature and is based on a sliding scale. It is best used to distinguish the relative placement of two or more models with respect to each other.

19. Heuristic Bias

Heuristic bias is based on the belief that humans use “mental short-cuts” for quick assessment and decision making. Through the use of heuristics, experts make decisions without detailed exploration and analysis of the problems space and all possible solutions. This allows for an acceptable although not necessarily optimal assessment of the situation or solution to an issue [STEI 98].

20. Human-Behavior Representation

A *human-behavior representation* (HBR) is “a model or simulation of any human function, any individual human, or any group or organization of humans.” [DEPA 01b] In this research, HBR will refer to the human cognitive process.

21. Human-Behavior Representation Knowledge Base

“The *HBR's knowledge base* contains the computer program that determines the HBR's response to the stimuli it receives from the simulated world. At a minimum, the knowledge base largely determines the HBR's cognitive behavior. It may also contribute to the manifestations of emotion upon behavior” [DEPA 01f].

22. Informational Bias

Informational or cognitive bias occurs when individuals use “intuitive strategies” to acquire and analysis information rather than using proven “optimal” methodologies. This results in the improper interpretation and presentation of data leading to non optimal solutions or improper conclusions. Sage describes twenty seven types of cognitive bias [SAGE 81].

23. Level

The assessment of behaviors is broken into three separate **levels** (e.g. task, scenario, and overall) which consist of sublevel assessments (e.g. subtasks, tasks, and scenarios, respectively). These create **level and sublevel pairings** (e.g. subtask to task, task to scenario, and scenario to overall).

24. Measure of Effectiveness (MOE)

A **measure of effectiveness** (MOE) is “a qualitative or quantitative measure of aggregate performance or a characteristic of a model, simulation or system that indicates the degree to which it performs the task or meets an operational objective or requirement under specified conditions” [DEPA 95].

25. Measure of Performance (MOP)

A **measure of performance** (MOP) is “the measure of how the system/individual performs its functions in a given environment (e.g., number of targets detected, reaction time, number of targets nominated, susceptibility of deception, task completion time). It is closely related to inherent parameters (physical and structural) but measures attributes of system behavior” [DEPA 97].

26. Model

A **model** is “a physical, mathematical, or otherwise logical representation of a system, entity, phenomenon, or process” [DEPA 01b].

27. Naturalistic Decision-Making

Naturalistic decision-making (NDM) is the study of how people use their experiences to make decisions in real-world situations. Its focus is on time-pressured decision-making processes used by experts when information is missing or ambiguous, goals are vague, and conditions are changing [KLEI 01].

28. Normative Bias

Normative bias is concerned with the interaction between individuals who provide information or skills to the community in order to resolve an issue or cultivate a conclusion [DUFF 93].

29. Overall

The *overall* assessment is the final judgment of the model/individual performance derived from a collection of scenarios. For this research, the overall assessment is how well the SME feels the individuals and leader performed their roles.

30. Participant/Rater

A *participant* or *rater* is an individual taking part in the experiments who performs an assessment of observed model/human behaviors. The participants in this research come from a pool of 182 Army and USMC officers enrolled in the Infantry Captains Carrier Course at Fort Benning, GA. This document refers to these individuals as subject-matter experts (SMEs).

31. Perception Bias

Perception bias is that which an expert brings to the process based on his education, training, real-world experiences, exposure to simulations, and organizational loyalties. These factors color the lenses of the SME's microscope or unduly focus the search area on certain aspects of a model's performance [PACE 02].

32. Performance Bias

Performance bias deals with the SME's ability to execute the validation process. This ability may be hampered by other demands on the SME's time, the inavailability of data, a low ability or desire to comply with specified validation procedures, or the expert's failure to understand the simulation [PACE 02].

33. Perspective Bias

Perspective bias occurs when a SME's fails to maintain focus on the intended purpose of the model. A SME may lose focus as he allows his real-world experiences to cloud his view on what the model should have the capability of doing [PACE 02].

34. Rating

An *assessment* or *rating* is the value (based on scale) an individual SME gives an observed model or human behavior.

35. Referent

“A codified body of knowledge about a thing being simulated” [HARM 98]. In the case of HBR and this research, this would consist of at least one of the six levels of correspondence. *Referent* is the best information we have about the simulated object’s functionality and performance. The referent provides the standards against which the results of models and simulations are compared, to assess the level of fidelity they are able to replicate [DEPA 00] [DEPA 01f].

36. Resolution

Different from fidelity, *resolution* is “the degree of detail used to represent aspects of the real world or a specified standard or referent by a model or simulation” [DEPA 01b]. Resolution often refers to the visual characteristics of a model.

37. Scale

A *scale* is a set of possible assessment responses SMEs can use to quantify the level of performance of the observed behavior. Three scales are used in this research. Scale 1 is a seven-point Likert scale, where a seven represents the SME’s highest confidence the model or individual performed to standard and one indicates the SME’s certainty that the model or individual failed to perform to standard. Scale 2 is a Go/No-Go scale where a Go indicates the SME’s belief the model or individual performed to standard and No-Go indicates the belief that the model or individual failed to perform to standard. Scale 3 is a five-point Likert scale where five represents the SME’s highest confidence the model or individual performed to standard and one indicates the SME’s utmost confidence the model or individual failed to perform the behavior to standard.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

- [AIR 01] Air Force Research Laboratory (AFRL) (01 Jun 01). *Agent-based Modeling and Behavior Representation (AMBR) Project*. [WWW Document]. <https://www.williams.af.mil/html/ambr.htm> (Viewed 21 May 2002).
- [AMSO 00] Army Modeling and Simulation Office (AMSO) (15 September 2000). "Planning Guidelines for Simulation and Modeling for Acquisition Requirements and Training" [WWW Document]. <http://www.amso.army.mil/smart/documents/guidelines/guidelines.doc> (Viewed 25 July 2002).
- [ARON 02] Aronson, W. S. (September 2002). *A Cognitive Task Analysis of Close Quarters Battle*. (Unpublished Master's Thesis). Monterey, CA: Computer Science Department, Naval Postgraduate School.
- [ARTP 01] *ARTEP 7-8-MTP: Mission Training Plan for the Infantry Rifle Platoon and Squad*. (Mission Training Plan) (2001). Washington, DC: Headquarters, Department of the Army.
- [BALC 97] Balci, O. (1997). "Verification, Validation and Accreditation of Simulation Models." In *Proceedings of the 1997 Winter Simulation Conference* (Atlanta, GA, Dec. 7-10). IEEE, Piscataway, NJ, pp 135-141.
- [BARN 93] Barnett, B. J., Perrin, B. M., & Walrath, L. D. (1993). *Bias in Human Decision-making for Tactical Decision-making Under Stress*. St. Louis, MO: McDonnell Douglas Corporation.
- [BIRT 96] Birta, L.G. & Özmırak, F.N. (January 1996). A Knowledge-Based Approach for the Validation of Simulation Models: The Foundation. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 6(1), 76-98. [WWW Document]. <http://doi.acm.org/10.1145/229493.229511> (Viewed 23 August 2002).
- [BLAC 03] Black, P. E. (2003, 06 January). *Halting Problem*. [WWW Document]. <http://www.nist.gov/dads/HTML/haltingProblem.html> (Viewed 27 May 2004).
- [BUST 02] Bustamante, L., Howe-Tennant, D., and Ramo, C. (2002). *The Behavioral Approach*. Tutorial for PSY 501: Advanced Educational Psychology. Cortland, NY: SUNY College. [WWW Document]. <http://facultyweb.cortland.edu/~ANDERSMD/BEH/BEHAVIOR.HTML> (Viewed 02 September 2002).

- [BUZI 00] Buziak, C. (2000). *The Role of Personality in Determining Variability in Evaluation Expertise*. Unpublished Master's Thesis, Naval Postgraduate School, Monterey, CA.
- [CARD 80] Card, S. K., Moran, T. P., & Newell, A. (1980). The Keystroke-Level Model for User Performance with Interactive Systems. *Communications of the ACM*, 23(7), 396-410.
- [CARD 83] Card, S. K., Moran, T. P., & Newell, A. (1983). *The Psychology of Human-Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- [CART 98a] Carter, T (March 1998). *2.8 Test and Evaluation*. Office of the Secretary of Defense/Director of Operational Test & Evaluation (OSD/DOT&E), [WWW Document]. <http://web2.deskbook.osd.mil/irp/28.asp?HEADER=NO&L2-8> (Viewed 02 July 2002).
- [CART 98b] Carter, T (March 1998). *2.8.1 Test and Evaluation Process*. Office of the Secretary of Defense/Director Of Operational Test & Evaluation (OSD/DOT&E), [WWW Document]. <http://web2.deskbook.osd.mil/irp/281.asp?HEADER=NO&L3-43> (Viewed 02 July 2002).
- [CASC 98] Cascio, W. F. (1998). *Applied Psychology in Human Resource Management* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
- [CAUG 95] Caughlin, D. (1995). *Verification, Validation and Accreditation (VV&A) of Models and Simulations through Reduced Order Metamodels*. Arlington, VA: Proceedings of the 27th Conference on Winter Simulation. pp. 1405-1412. [WWW Document]. <http://doi.acm.org/10.1145/224401.224832> (Viewed 23 August 2002).
- [CHAR 02] Charlton, S. G., & O'Brien, T. G. (Eds.). (2002). *Handbook of Human Factors Testing and Evaluation* (2nd ed.). Mahwah, N.J.: Lawrence Erlbaum Associates, Inc.
- [CHON 01] Chong, R. (2001). *Low-Level Behavioral Modeling and the HLA: EPIC-Soar Model of an Enroute Air-Traffic Control Task*. Norfolk, VA: Proceedings of the Tenth Conference on Computer Generated Forces and Behavioral Representation.
- [COHE 93] Cohen, M. S. (1993). The Naturalistic Basis of Decision Biases. In G. A. Klein, J. Orasanu, R. Calderwood & C. E. Zsombok (Eds.), *Decision-making in Action: Models and Methods* (pp. 51-99). Norwood, NJ: Ablex Publishing.

- [COHE 97] Cohen, M. S., Freeman, J. T., & Thompson, B. B. (1997). Training the Naturalistic Decision Maker. In C. E. Zsombok & G. A. Klein (Eds.), *Naturalistic Decision-Making: Expertise-Research and Applications* (pp. 257-268). Mahwah, NJ: Lawrence Erlbaum Associates.
- [COST 00] Costa, P. T., & McCrae, R. R. (2000). *NEO Software System™ for Windows Manual*. Odessa, FL: PAR Psychological Assessment Resources, Inc.
- [COST 03] Costa, P. T., & McCrae, R. R. (2003). *NEO Five-Factor Inventory*. Lutz, FL: PAR Psychological Assessment Resources, Inc.
- [COST 92] Costa, P. T., & McCrae, R. R. (1992). *NEO PI-R Professional Manual*. Odessa, FL: PAR Psychological Assessment Resources, Inc.
- [DANN 02] Dannemiller, J.L. (Ed) (January 2002). Journal Description. *Developmental Psychology*. Washington, DC: American Psychological Association [WWW Document]. <http://www.apa.org/journals/dev/description.html> (Viewed 02 September 2002).
- [DEAN 95] Dean, R., Allen, J., and Aloimonos, Y. (1995) *Artificial Intelligence; Theory and Practice*. Menlo Park, CA: Addison-Wesley Publishing Company.
- [DEPA 00a] Department of Defense Modeling and Simulation Office (DMSO) (30 November 2000). *Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide (RPG) Special Topic - Subject Matter Experts and VV&A*. [WWW Document]. http://vva.dmsomil/Special_Topics/SME/sme-pr.PDF (Viewed 23 January 2003).
- [DEPA 00b] Department of Defense Modeling and Simulation Office (DMSO) (30 November 2000). *Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide (RPG) Special Topic – Validation*. [WWW Document]. http://www.msiac.dmsomil/vva/Special_topics/Validation/Validation.htm (Viewed 10 July 2002).
- [DEPA 01] Department of Defense Directive (04 January 2001). *(DoDD) 5000.1: The Defense Acquisition System*. [WWW Document]. http://web2.deskbook.osd.mil/htmlfiles/rlframe/REFLIB_Frame.asp?TOC=/htmlfiles/TOC/061ddtoc.asp?sNode=L4-6&Exp=N&Doc=/reflib/mdod/061dd/061ddd.doc.htm&BMK=T16 (Viewed 02 July 2002).

- [DEPA 01a] Department of Defense Instruction (05 October 2001). (*DoD Instruction 5000.61: DoD Modeling and Simulation (M&S) Verification, Validation and Accreditation (VV&A)*). [WWW Document]. https://www.dmsso.mil/public/library/projects/vva/products/dodi_5000.61_recoordination_document_5_oct.pdf (Viewed 27 June 2002).
- [DEPA 01b] Department of Defense Modeling and Simulation Office (DMSO) (15 October 2001). *Verification, Validation and Accreditation Glossary*. [WWW Document]. <https://www.dmsso.mil/public/library/projects/vva/glossary.pdf> (Viewed 26 July, 2002).
- [DEPA 01c] Department of Defense Modeling and Simulation Office (DMSO) (15 August 2001). *Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide (RPG) Reference Document - A Practitioner's Perspective on Simulation Validation and Conceptual Model Development and Validation*. [WWW Document]. http://www.msiac.dmsso.mil/vva/ref_docs/val_lawref/default.htm#toc1 (Viewed 10 July 2002).
- [DEPA 01d] Department of Defense Modeling and Simulation Office (DMSO) (15 August 2001). *Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide (RPG) Reference Document - Human Behavior Representation (HBR) Literature Review*. [WWW Document]. http://www.msiac.dmsso.mil/vva/Ref_Docs/HBR/beh-ref-pr.PDF (Viewed 27 June 2002).
- [DEPA 01e] Department of Defense Modeling and Simulation Office (DMSO) (15 August 2001). *Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide (RPG) Reference Document - Key Concepts of VV&A*. [WWW Document]. <http://www.msiac.dmsso.mil/vva/Key/key.pr.pdf> (Viewed 27 June 2002).
- [DEPA 01f] Department of Defense Modeling and Simulation Office (DMSO) (25 September 2001). *Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide (RPG) Special Topic - Validation of Human Behavior Representations*. [WWW Document]. http://www.msiac.dmsso.mil/vva/Special_topics/hbr-Validation/default.htm (Viewed 27 June 2002).
- [DEPA 01g] Department of the Army Field Manual (01 March 2001). *FM 7-8: Infantry Rifle Platoon and Squad*. (2001). Washington, DC: Headquarters, Department of the Army.

- [DEPA 94] Department of Defense Directive (04 January 1994). “(DoDD) 5000.59: DoD Modeling and Simulation (M&S) Management”, [WWW Document].
http://www.ailtso.com/simval/Documents/5000.59/dod_directive_5000.59.html (Viewed 26 June 2002).
- [DEPA 95] Department of Defense (October 1995). “DoD 5000.59-P: Modeling and Simulation (M&S) Master Plan”, [WWW Document].
<http://www.dmsso.mil> (Viewed 26 June 2002).
- [DEPA 97] Department of Defense (December 1997). “DoD 5000.59-M: Modeling and Simulation (M&S) Glossary.”
- [DEPA 97a] Department of the Army Field Manual (30 September 1997). “FM 101-5-1/MCRP 5-2A: Operations Terms and Graphics.” Washington, DC: Headquarters, Department of the Army.
- [DRUK 88] Druckman, D., & Swets, J. A. (Eds.). (1988). *Enhancing Human Performance: Issues, Theories, and Techniques*. Washington, DC: National Academy Press.
- [DUFF 93] Duffy, L. (1993). Team Decision-Making Biases: An Information-Processing Perspective. In G. A. Klein, J. Orasanu, R. Calderwood & C. E. Zsombok (Eds.), *Decision-making in Action: Models and Methods* (pp. 346-359). Norwood, NJ: Ablex Publishing.
- [ENCY 02] *Encyclopedia Britannica* (04 February 2002). Encyclopedia Britannica, Inc. [Website] <http://www.britannica.com/> (Viewed 02 September 2002).
- [ERIC 95] Ericsson, K. A., & Kintsch, W. (1995). Long-Term Working Memory. *Psychological Review*, 102(2), 211-245.
- [FREE 99] Freedman, A. (1999). *The Computer Desktop Encyclopedia* (CD Rom - Ver 12.1). Point Pleasant, PA: Computer Language Company Inc.
- [GALE 01] Gale Group (2001). *The Gale Encyclopedia of Psychology*, (2nd Ed). Detroit, MI: Gale Group. [WWW Document].
http://www.findarticles.com/cf_0/g2699/0000/2699000080/p1/article.jhtml (Viewed 12 December 2002).
- [GALL 03] Galligan, D. P., Anderson, M. A., & Lauren, M. K. (2003). *MANA, Map Aware Non-uniform Automata, Version 3.0, Users Manual (Dr.aft)*. Unpublished manuscript.

- [GAWR 00] Gawron, V. J. (2000). *Human Performance Measures Handbook*. Mahwah, N.J.: Lawrence Erlbaum Associates.
- [GILO 02] Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and Biases: The Psychology of Intuitive Judgement*. Cambridge, NY: Cambridge University Press.
- [GOER 02] Goerger, S. R. (22 August 2002). *Validation and Evaluation of Cognitive Architectures Using an Emergent Combat Model*. Presentation at the MOVES Open House, Monterey, CA.
- [GOER 03] Goerger, S. R. (2003, 20 - 24 July). *Validating Human Behavioral Models for Combat Simulations Using Techniques for the Evaluation of Human Performance*. Paper presented at the 2003 Summer Computer Simulation Conference, Montreal, Quebec, Canada.
- [GONZ 98] Gonzalez, A.J. and Murillo, M. (04 Dec 98). *Validation of Human Behavior Models*. (Paper Number 99S-SIW-010) Simulation Interoperability Standards Organization's (SISO) 1999 Spring Simulation Interoperability Workshop (SIW). [WWW Document]. http://www.sisostds.org/doclib/doclib.cfm?SISO_FID_2442 (Viewed 24 July 2002).
- [GRAY 00] Gray, W.D. (July 2000). *Summary of the AMBR Expert Panel Report*. [WWW Document] https://www.williams.af.mil/AMBR/AMBR1_Gray.ppt (Viewed 21 May 2002).
- [GROS 99] Grossman, D. (1999). *Behavioral Psychology*. Jonesboro, AR: Killology Research Group. [WWW Document]. http://www.killology.com/article_behavioral.htm (Viewed 02 September 2002).
- [HALE 02] Hale, J. (2002). *Performance Based Evaluation: Tools and Techniques to Measure the Impact of Training*. San Francisco, CA: Jossey-Bass.
- [HARM 03] Harmon, S. Y. (04 August 2003). *Levels of VV&A*. Presentation at the Navy Modeling and Simulation Management Office (NAVMSMO) Verification, Validation & Accreditation (VV&A) Technical Working Group (TWG) Workshop #14, Naval Postgraduate School, Monterey, CA.
- [HARM 98] Harmon, S.Y. (Ed.). (16 December 1998). *Fidelity ISG Glossary*. Ver 3.0. Simulation Interoperability Standards Organization (SISO), Fidelity Implementation Study Group (ISG), [WWW Document]. http://www.sisostds.org/doclib/doclib.cfm?SISO_RID_1000789 (Viewed 02 July 2002).

- [HARV 01] Harvey, C.M. (2001). *Cognitive Task Analysis*. (briefing slides) Dayton, OH: Dept. of Biomedical, Industrial, and Human Factors Engineering, Wright State University [WWW Document]. <http://champ.cs.wright.edu/ai/hfe733-01/cta.pdf> (Viewed 04 September 2002).
- [HODG 92] Hodges, J. and Dewar, J. (1992). *Rand Corporation Report R-4114-RC/AF; Is It Your Model Talking? A Framework for Model Validation*. Santa Monica, CA: Rand Corporation.
- [HOFF 03] Hoffman, C. W. D. (2003). *A Structured Engineering Process Standard for CGF Behavioral Development*. (Unpublished Presentation) Orlando, FL: OneSAF.
- [HOLL 95] Holland, J.H. (1995). *Hidden Order: How applications Build Complexity*. Cambridge, MA: Perseus Books.
- [HUGH 97] Hughes, W.P. (editor) (1997). *Military Modeling for Decision-making* (3rd Ed). Alexandria, VA: Military Operations Research Society.
- [HUTC 96a] Hutchins, S. G., Kelly, R. T., & Morrison, J. G. (1996, June 25-28). *Decision Support for Tactical Decision-making Under Stress*. Paper presented at the 40th Human Factors and Ergonomics Society Annual Meeting, Monterey, CA.
- [HUTC 96b] Hutchins, S. G., Kelly, R. T., & Morrison, J. G. (1996, June 25-28). *Principles for Aiding Complex Military Decision-making*. Paper presented at the 40th Human Factors and Ergonomics Society Annual Meeting, Monterey, CA.
- [ILAC 97] Ilachinski, A. (August 1997). *CRM 97-61.10: Irreducible Semi-Autonomous Adaptive Combat (ISAAC): An Artificial-Life Approach to Land Warfare (U)*. Alexandria, VA: Center for Naval Analyses.
- [INTR 02] Introduction to Joint Combat Modeling (OA/MV4655) (July 2002). *Combat Modeling Overview*. Class Notes. Monterey, CA: Naval Postgraduate School Operations Research Department.
- [JONA 99] Jonassen, D.H., Tessmer, M., and Hannum, W.H. (1999). *Task Analysis Methods for Instructional Design*. Mahwah, NJ: Lawrence Erlbaum Associates.
- [JONE 99] Jones, R.M. Laird, J. E., Nielsen, P. E., Coulter, K. J., Kenny, P., & Koss, F. V. (1999). Automated Intelligent Pilots for Combat Flight Simulation. *AI Magazine*. Spring, 1999.

- [KAHN 82] Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge, NY: Cambridge University Press.
- [KING 80] King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in a Multi-Dimensional Forced-Choice Performance Evaluation Scale. *Journal of Applied Psychology*, 65, 507-516.
- [KLEI 00] Klein, G. A. (2000). Cognitive Task Analysis of Teams. In J. M. Schraagen, S. F. Chipman & V. L. Shalin (Eds.), *Cognitive Task Analysis* (pp. 417-429). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- [KLEI 01] Klein, G. (2001). *Sources of Power: How People Make Decisions*. Cambridge, MA: The MIT Press.
- [KLEI 97] Klein, G. A. (1997). An Overview of Naturalistic Decision-Making Applications. In C. E. Zsombok & G. A. Klein (Eds.), *Naturalistic Decision-Making: Expertise-Research and Applications* (pp. 49-59). Mahwah, NJ: Lawrence Erlbaum Associates.
- [LAIR 00] Laird, J.E. and van Lent, M. (2000). *Human-Level AI's Killer Application: Interactive Computer Games*. Proc. American Association of Artificial Intelligence (AAAI) 2000. Cambridge, MA: AAAI Press / The MIT Press: pp 1171-1178.
- [LAUR 02] Lauren, M. K., & Stephen, R. T. (2002). Map-aware Non-uniform Automata (MANA) - A New Zealand Approach to Scenario Modelling. *Journal of Battlefield Technology*, 5(1).
- [LEBI 01] Lebiere, C., Anderson, J.R., and Bothel, D. (2001). *Multi-Tasking and Cognitive Workload in an ACT-R Model of a Simplified Air Traffic Control Task*. (Paper Number 10th-CGF-071) Norfolk, VA: Proceedings of the Tenth Conference on Computer Generated Forces and Behavioral Representation.
- [LEWI 02] Lewis, T.G., Zyda, M., and Hiles, J.E (2002). *Proposal to Establish a Center for Study of Potential Outcomes*. (Unpublished Technical Report) Monterey, CA: Naval Postgraduate School, Modeling of Virtual Environments and Simulations (MOVES) Institute.
- [LIPS 97a] Lipshitz, R. (1997). Naturalistic Decision-Making Perspectives on Decision Errors. In C. E. Zsombok & G. A. Klein (Eds.), *Naturalistic Decision-Making: Expertise-Research and Applications* (pp. 151-162). Mahwah, NJ: Lawrence Erlbaum Associates.

- [LIPS 97b] Lipshitz, R., & Shaul, O. B. (1997). Schemata and Mental Models in Recognized-Primed Decision-making. In C. E. Zsombok & G. A. Klein (Eds.), *Naturalistic Decision-Making: Expertise-Research and Applications* (pp. 293-204). Mahwah, NJ: Lawrence Erlbaum Associates.
- [MALL 88] Mallery, J. C. (1988, 28 March - 03 April). *Thinking About Foreign Policy: Finding an Appropriate Role for Artificially Intelligent Computers*. Paper presented at the The 1988 Annual Meeting of the International Studies Association, Adam's Mark Hotel, St. Louis, MO.
- [MATT 00] Matthews, G, Davies, D.R., Westerman, S.J., and Stammers, R.B. (2000). *Human Performance: Cognition, Stress and Individual Differences*. Philadelphia, PA: Psychology Press, Taylor & Francis Group.
- [MERR 02] *Merriam-Webster's Collegiate Dictionary*. [WWW Document]. <http://www.m-w.com/cgi-bin/dictionary> (Viewed 18 July 2002).
- [MERR 03] *Merriam-Webster's Collegiate Dictionary*. [WWW Document]. <http://www.m-w.com/cgi-bin/dictionary> (Viewed 09 December 2003).
- [MILI 97] Militello, L (ed) (1997). *ACTA: Applied Cognitive Task Analysis Instructional Software*. (CD Rom). Fairborn, OH: Klein Associates Inc.
- [MILL 04] Miller, N. L., & Shattuck, L. G. (2004). *Human Performance; Report on DARPA's Future Combat System Command and Control, Experiment 4a, 28 September - 2 October 2003*: Defense Advanced Research Projects Agency (DARPA).
- [MILL 56] Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review*, 63, 81-97.
- [MILL 97] Miller, T. E., & Woods, D. D. (1997). Key Issues for Naturalistic Decision-Making Researchers in System Design. In C. E. Zsombok & G. A. Klein (Eds.), *Naturalistic Decision-Making: Expertise-Research and Applications* (pp. 141-150). Mahwah, NJ: Lawrence Erlbaum Associates.
- [MITC 97] Mitchell, C. M., Morris, J. G., Ockerman, J. J., & Potter, W. J. (1997). Recognition-Primed Decision-making as a Technique to Support Reuse in Software Design. In C. E. Zsombok & G. A. Klein (Eds.), *Naturalistic Decision-Making: Expertise-Research and Applications* (pp. 305-318). Mahwah, NJ: Lawrence Erlbaum Associates.

- [NOFI 00] Nofi, A.A. (November 2000). *CRM D0002895.A1/Final: Defining and Measuring Shared Situational Awareness*. Alexandria, VA: Center for Naval Analyses. [WWW Document]. <http://www.cna.org/newsevents/images/crmd2895final.pdf> (Viewed 29 July 2002).
- [ONES 03] *OneSAF Objective System*. (2003, 11 November). Retrieved 09 December, 2003, from <http://www.onesaf.org/public1saf.html>
- [OSBO 02] Osborne, B.A. (September 2002). *An Agent-Based Architecture for Generating Interactive Stories*. (Unpublished Dissertation). Monterey, CA: Naval Postgraduate School, Computer Science Department.
- [PACE 02] Pace, D. K., & Sheehan, J. (2002, 22-24 October). *Subject Matter Expert (SME)/Peer Use in M&S V&V*. Paper presented at the Foundations '02, a Workshop on Model and Simulation Verification and Validation for the 21st Century, Kossiakoff Conference & Education Center, Johns Hopkins University Applied Physics Laboratory, Laurel, MD.
- [PACE 04] Pace, D. K. (Dale.Pace@jhuapl.edu) (08 March 2004). RE: Validation Plan for Dissertation Research. Email to S. R. Goerger (srgoerge@nps.navy.mil).
- [PERR 93] Perrin, B. M., Barnett, B. J., & Walrath, L. D. (1993). *Techniques to Reduce Bias in Human Decision-making for Tactical Decision-making Under Stress (Tasks 2 & 3)*. St. Louis, MO: McDonnell Douglas Corporation.
- [PETE 85] Petersen, R. G. (1985). *Design and Analysis of Experiments* (Vol. 66). New York, NY: Marcel Dekker.
- [PEW 98] Pew, R. W., & Mavor, A. S. (Eds.). (1998). *Modeling Human and Organizational Behavior: Application to Military Simulations*. National Academy Press, Washington, DC.
- [PIAN 01] Pianta, D. (2001). *Design Methods Fact Sheet: Measures of Effectiveness (MOE) and Measures of Performance (MOP)*. Retrieved 27 August, 2003, from http://www.catalyst.uq.edu.au/designsurfer/MoE_MoP.pdf
- [POND 02] Pounder, J. S., & Mun, T. (28 February 2002). *A Behaviourally Anchored Rating Scales Approach to Institutional Self Assessment in Higher Education*. Retrieved 01 April 2003, [WWW Document]. <http://www.ugc.edu.hk/english/documents/papers/pounder.html> (Viewed 01 April 2003).

- [PREN 01] Prensky, M. (2001). Digital Natives, Digital Immigrants. *On the Horizon*, 9(5), 1-6.
- [PROJ 01] *Project Albert Fact Description* (2001). [WWW Document]. <http://www.mcwl.quantico.usmc.mil/divisions/albert/index.asp> (Viewed 01 April 2004).
- [PROJ 02] *Project Albert Fact Sheet* (10 December 2002). [WWW Document]. http://www.mcwl.quantico.usmc.mil/fact_sheets/fs/Pro%20Albert%2007_31_03.pdf (Viewed 01 April 2004).
- [RALS 00] Ralston, A., Reilly, E.D. and Hemmendinger, D. (2000). *Encyclopedia of Computer Science* (4th ed). New York, NY: Grove's Dictionaries, Inc.
- [RODD 00] Roddy, K. and Dixon, M. (September 2000). *Modeling Human and Organizational Behavior using a Relation-Centric Multi-Agent System Design Paradigm*. (Unpublished Master's Thesis) Monterey, CA: Naval Postgraduate School, MOVES Academic Group.
- [RUSS 95] Russell, S. & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ: Prentice Hall.
- [SAGE 81] Sage, A. P. (1981). Behavioral and Organizational Considerations in the Design of Information Systems and Processes for Planning and Decision Support. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-11(9), 640-678.
- [SCHE 78] Schelling, T. C. (1978). *Micromotives and Macrobehavior*. New York, NY: W.W. Norton & Co.
- [SHAT 03] Shattuck, L. G., & Miller, N. L. (2003). Situation Awareness in the Objective Force: Joining Technology and Cognition in Complex Systems (Draft). 14.
- [SIMU 98] Simulation Interoperability Standards Organization (SISO) Fidelity Implementation Study Group (ISG) (December 1998). *Fidelity ISG Glossary*, V 3.0, [WWW Document]. http://www.sisostds.org/doclib/doclib.cfm?SISO_RID_1000789 (Viewed 24 July 2002).
- [SMIT 63] Smith, P. C., & Kendall, L. M. (1963). Retranslation of Expectations: An Approach to the Construction of Unambiguous Anchors for Rating Scales. *Journal of Applied Psychology*, 47(149-155).

- [SOLD 03] *Soldier's Manual of Common Tasks Skill Level 2, 3, and 4.* (Soldier Training Publication) (2003). Washington, DC: Headquarters, Department of the Army.
- [SOLS 01] Solso, R.L. (2001). *Cognitive Psychology* (6th ed.). Needham Heights, MA: Allyn & Bacon.
- [STAT 03] Statkus, M. J., Sampson, J. B., & Woods, R. J. (2003). *Human Science/Modeling and Analysis Data Project: Situation Awareness Effects on Troop Movement and Decision-making Data Collection Effort, 21 October Through 1 November 2002* (Technical Report No. NATICK/TR-03/033L). Natick, MA: U.S. Army Soldier & Biological Chemical Command, Natick Soldier Center.
- [STEI 98] Stein, R., & Stein, M. (1998). Sources of Bias and Inaccuracy in the Development of a Best Estimate. *Casualty Actuarial Society Forum*, 45.
- [STUF 02] Stufflebeam, D. L. (19 November 2002). *Guidance for Choosing and Applying Evaluation Checklists*, [WWW Document]. <http://www.wmich.edu/evalctr/checklists/checklistorganizer.htm> (Viewed 12 December 2003)
- [TOLK 02] Tolk, A. (2002, 10-11 December). *Human Behaviour Representation - Recent Developments*. Paper presented at the NATO Research & Technology Organization (RTO); Studies, Analyses and Simulation Panel (SAS); Lecture Series on "Simulation of and for Military Decision-making", The Hague, Netherlands.
- [TURN 91] Turner, R. P. and Gibilisco, S. (Eds.) (1991). *Illustrated Dictionary of Electronics* (5th ed). Blue Summit, PA: TAB Professional & Reference Books.
- [TVER 71] Tversky, A., & Kahneman, D. (1971). Belief in the Law of Small Numbers. *Psychological Bulletin*, 76, 105-110.
- [TVER 74] Tversky, A., & Kahneman, D. (1974). Judgement Under Uncertainty: Heuristics and Biases. *Science*, 185, 1124-1130.
- [TZIN 00] Tziner, A., Joanis, C., & Murphy, K. R. (2000). A Comparison of Three Methods of Performance Appraisal with Regard to Goal Properties, Goal Perception and Ratee Satisfaction. *Group and Organization Management*, 25(2), 175-190.
- [US L 99] *US Land Warfare Systems*. (1999). Retrieved 06 January, 2004, from <http://www.fas.org/man/dod-101/sys/land/>

- [WELL 03] Wellbrink, J. (2003). *A Reduced Human Performance Model for Exploring Unintended Consequences and Potential Outcomes*. Unpublished Dissertation, Naval Postgraduate School, Monterey, CA.
- [WEST 02] West, M.J. (ed) (June 2002). Journal Description. *Journal of Comparative Psychology*. Washington, DC: American Psychological Association [WWW Document]. <http://www.apa.org/journals/com/description.html> (Viewed 02 September 2002).
- [WILS 99] Wilson, R.A., & Keil F.C. (1999). *The MIT Encyclopedia of the Cognitive Sciences* (CD Rom). Cambridge, MA: MIT Press.
- [WRAY 92] Wray, R., Chong, R. Phillips, J., Rogers, S., and Walsh, B. (1992). *A Survey of Cognitive and Agent Architectures*. Ann Arbor, MIP: Department of Electrical Engineering and Computer Science, University of Michigan. [WWW Document]. <http://ai.eecs.umich.edu/cogarch0/index.html> (Viewed 06 September 2002).
- [ZACK 01] Zackery, W. (2001). *Developing a Multi-task Cognitive Agent Using the COGNET/iGEN Integrative Architecture*. Norfolk, VA: Proceedings of the Tenth Conference on Computer Generated Forces and Behavioral Representation.
- [ZSAM 97] Zsombok, C. E. (1997). Naturalistic Decision-Making: Where are We Now? In C. E. Zsombok & G. A. Klein (Eds.), *Naturalistic Decision-Making: Expertise-Research and Applications* (pp. 3-16). Mahwah, NJ: Lawrence Erlbaum Associates.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
Ft. Belvoir, VA
2. Dudley Knox Library
Naval Postgraduate School
Monterey, CA
3. Dr. Michael Zyda
Director, MOVES Institute
Naval Postgraduate School
Monterey, CA
4. Dr. Rudy Darken
Department of Computer Science
Naval Postgraduate School
Monterey, CA
5. COL Michael McGinnis
Systems Engineering Department
United States Military Academy
West Point, NY
6. Dr. Chris Darken
Department of Computer Science
Naval Postgraduate School
Monterey, CA
7. Dr. Nita Miller
Department of Operations Research
Naval Postgraduate School
Monterey, CA
8. Sue Hutchins
Department of Information Science
Naval Postgraduate School
Monterey, CA
9. Army Modeling and Simulations Office
HQDA, DCS G3 (DAMO-ZS)
400 Army Pentagon
Washington, DC

10. Navy Modeling and Simulation Management Office (N61M)
2000 Navy Pentagon
Washington, DC